nature

More good news from Moscow

Hoop-la arising from last week's summit meeting in Moscow should not elsewhere hide even more significant developments, of which the Soviet ambition to join GATT is the most striking.

THE meeting between President Ronald Reagan of the United States and Mr Mikhail Gorbachev, the secretary-general of the Soviet Communist Party, may have filled the newspapers in the Soviet Union and the West last week, but what may be the most telling of a flood of announcements from Moscow has been given less attention than it deserves. By all accounts and, specifically, according to the Soviet Commission on Foreign Economic Relations, the Soviet Union intends to apply for membership of the General Agreement on Tariffs and Trade, termed GATT, two years from now and even has an ambition to make the rouble freely convertible into other currencies, perhaps as soon as the end of the century. On the face of things, these good intentions do not match the other developments last week in Moscow — not least, Reagan's belated conversion to the view that Gorbachev may be the best thing to have happened to the Soviet Union in 70 years, but also the ratification (delayed as well) of the INF treaty. Yet, in the long run, the Soviet Union's wish to join the now half-global economic system will be more important, not merely for Soviet citizens but for the rest of us.

This is why. Convertible currencies and GATT rules have been the principal driving forces of the remarkable expansion in the past several decades of world trade, whose annual volume has grown even faster than the domestic economy of Japan. That trend, in its quiet way, has done as much to change the world as any of the other momentous tendencies since the Second World War — the threat that nuclear weapons might be used, the threat of climatic change not excepted. What the growth of world trade has accomplished is a decisive (if needlessly slow) improvement in the economic division of labour and, as a consequence, the astonishing enrichment of the people of the industrialized world and a general realization of their interdependence. But the isolation of the Soviet Union and its neighbours from this system has meant that the division of labour has not been as effective as it might have been, that both halves of the world's economy (but the Soviet half especially) have been needlesly impoverished and that the economic disparities have engendered political rivalry and tension. That is the sense in which the new Soviet ambition may be more important even than the INF treaty.

Ambition

But is the ambition attainable? While the decision to seek to belong to the Western trading system is by no means the most breathtaking of the changes to have been brought about by Moscow in the past two years, it will be a monumental break with the past, during which Soviet spokesmen have diligently branded the West's trading arrangements as gigantic cartels. It will be easier to tell which way the wind is blowing after the party conference arranged in Moscow for the 27 June.

But there are also technical, economic and social obstacles to be overcome. Joining GATT will require that the Soviet Union should have gone a long way to rationalize its domestic economy, charging realistic prices for the goods its people buy. There will be endless popular protests if the universal subway fare is increased, but there will also be heart-searching among the party's old guard about the compatibility of such realism with the theoretical foundations of the Soviet state. Certainly it will need a remarkable conjuring trick to make it all happen in the two years the ministry of overseas economic affairs is talking of. Making the rouble convertible is an even greater task, requiring nothing less than making a substantial slice of Soviet industry competitive with that of the West. In the Soviet Union and elsewhere, people will be keeping their fingers crossed.

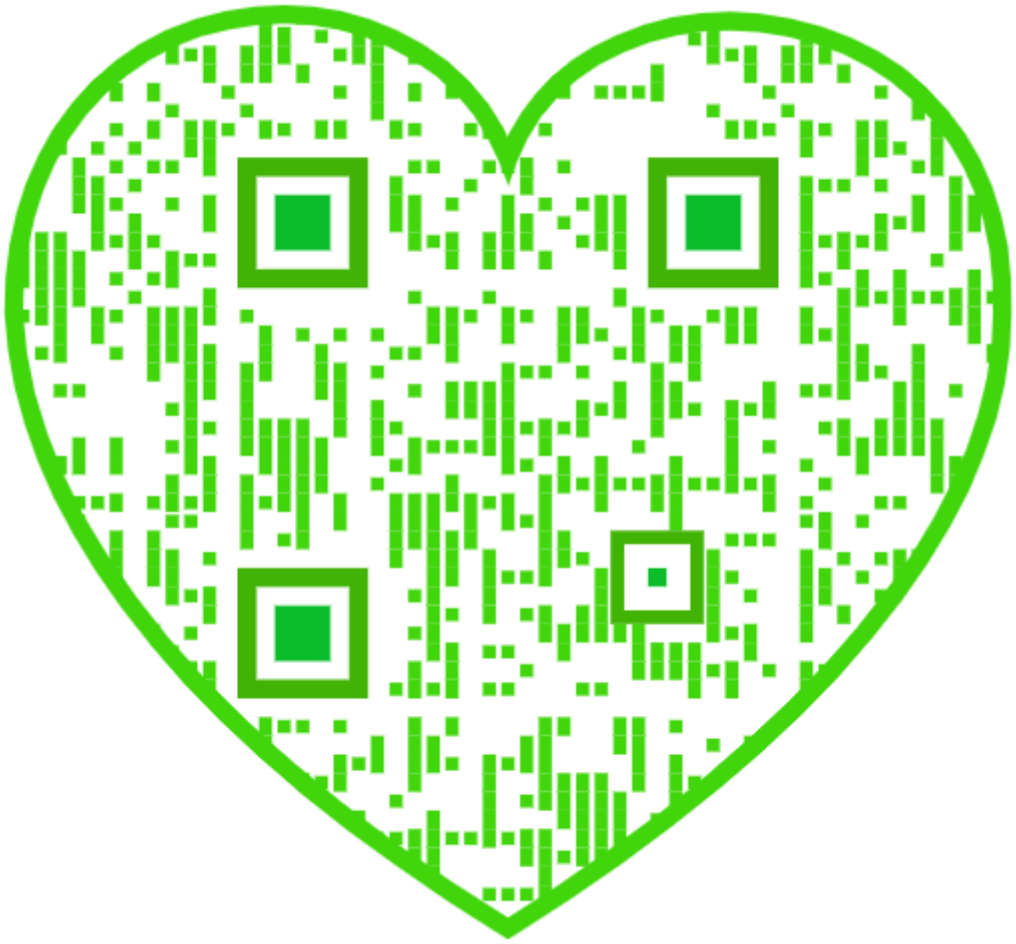
Prizes

Yet foreign trade will not be at the head of the agenda for the conference at the end of this month. Indeed, the ambition to join the rest of the world in trade is itself dependent on the outcome of more immediate decisions that will be needed — not merely the decentralization of the political system, but how to reorganize Soviet industry and to couple it more effectively to the huge but ineffectual research enterprise. But that will bring another prize for the West (not to mention the Soviet Union): a more fruitful relationship between two productive research communities separated from each other for far too long by external constraints. Those who marvel that the Englishman Humphry Davy should have been able to make an extended carriage ride in France at the height of the Napoleonic wars will not be alone in asking what the benefits of genuinely free exchange between the West and the Soviet Union would bring. Last week's summiteers offered general encouragement, but the world is waiting to see the colour of their money.

To some extent, the issue of free exchange is clouded by the separate issue of those who wish to leave the Soviet Union, but who are still denied exit visas. This was predictably one of the contentious issues at least week's summit meeting. But those in the West who champion the cause of the refusniks (many of whom are denied jobs as well as visas and many of whom are scientists) may, for all kinds of humanitarian reasons, be flogging last decade's horse. If indeed the social upheaval under way in the Soviet Union turns out to be as profound as promised, may it not be that the interests of refusniks would be better served by effective restraints of racial and other prejudice than by exit visas, especially at such a time?

Much the same is true of the failure last week to reach an understanding on the promised treaty to reduce superpower strategic missiles by a half. Already (see *Nature* **333**, 284; 1988) there are good grounds for believing that a fully verifiable treaty will be too complicated to administer.

The fact that the two foreign ministers in Moscow last week found themselves signing (perhaps only for cosmetic purposes) an agreement to regulate the carrying capacity of bomber aircraft for cruise missiles is a sign of how the details of the proposed treaty are in danger of getting out of hand. The process should in any case be simplified if the next US administration turns its back on the Strategic Defense Initiative (SDI), or merely lets it wither away. (Mr Michael Dukakis, likely to have been confirmed this week as the probable Democratic candidate for president, calls SDI a "fraud"; Mr George Bush, his almost certain opponent, is lukewarm about this as about most things.) Why not give the small armies of negotiators something more tangible into which to sink their teeth in the coming election months? □



New ways with drugs

Legalizing hard drugs, but controlling them, could lessen the social evils of drug abuse.

NARCOTIC drugs, by general consent, constitute an unparalleled social evil, but nobody can decide what should be done about them (and it). In the United States, where politicians during the election months ahead will be competing for the sympathy of anxious parents and social workers by advertising their willingness to deal firmly with drug traffickers, the issue has now been further dramatized (and confused) by the inevitable rapid spread of AIDS among those injecting heroin intravenously.

One measure of the common despair over drugs is the proposal that the US armed services should stand shoulder-to-shoulder with the US Customs Service in an attempt to keep out drugs by force — interdiction is the word. Continuing alertness is no doubt essential, but surely a problem as serious as that now identified in the United States deserves more radical remedies. Is there a chance that the hapless intravenous drug users infected with the AIDS virus may provide the spur?

Sadly, the elements of the problem of drug-taking in prosperous societies are now all too familiar. On the supply side, drugs have been for decades a continual source of violence and corruption. Over the years, all kinds of stratagems have been tried, and have failed, to interrupt supplies at their source, in South-East Asia, Turkey and Latin America, for example: paying one farmer not to grow poppies may merely provide an incentive for his fellows to follow the same route. Interdiction, subtle or by armed force, no doubt serves as a deterrent for some who traffic in drugs, but has the effect of maintaining the street price of what is sold, increasing the financial burdens lying on the drug users. All this is well-known, as the piteous consequences of addiction, social and pharmacological. What can be done?

To note that much of what is now done by way of remedy is paradoxical is not, by that means, to complain; the difficulties are too great, and the consequences of failure too serious, for that. But there is a curious irony in the practice of the health authorities in many large cities of providing intravenous drug users with an ample supply of hypodermic needles so as to abate the spread of AIDS. The practice, of course, is a sensible means of protecting drug users from an even greater catastrophe, yet implicitly it is tantamount to condoning criminal activity. And a year after the clamour in the United States for the random testing of corporate employees for pharmacological signs of drug usage, it is still not clear whether the practice does more good (deterrence) than harm (ostracism for those in need of social help). Nobody would pretend that there can be a simple solution, but disquiet about drugs does require that elements of a remedy should have consequences that are predictable.

That is one reason for listening attentively to the case for going some way towards legalizing drugs, even heroin and cocaine. The principle would be essentially that followed in the care of the handful of heroin addicts in Britain until the 1960s — declared addicts would be provided with supplies from official sources at a cost they could afford. The obvious and immediate benefit is that the huge profits now obtained from the drug trade would be quickly and drastically curtailed, as would be the steady stream of drug-related crime. A further benefit would be that drug-users would at least be accessible to persuasion that a course of treatment would be worthwhile. The obvious difficulty is that societies would be seen to condone practices now firmly classified as criminal. Theoretically, at least, there is also a danger that a supply of cheap drugs would increase the number of those dependent on them, especially if the system were so loosely administered that supplies intended for declared drug-users were diverted onto a black market.

While it would be rash to claim that such a system could be made to function in present circumstances, the chance that it might cries out for serious study. No doubt it would ironically emerge that state narcotics monopolies would have their work cut out to wean a substantial proportion of drug users away from their present sources of supply (which would, of course, remain illegal). Drug users would not willingly declare themselves, and might fear for the long-term security of their supplies or, more immediately, compulsory cold turkey. An essential component of such an arrangement would be legislation requiring that declared users would not be discriminated against in employment except to the extent that their performance is marred by drugs. Another is an understanding there will be funds for the generous support of drug-treatment clinics. If drugs are as much of a menace as they appear, can it seriously be said that the cost of cure is too high? □

We wuz robbed

Blood is now being spilled differently in British universities, making institutions lop-sided.

SOONER or later, it had to happen that some long-suffering head of a British university would break with the conventions of academic civility and complain publicly of having been robbed. Dr Eric Ash, the rector (equivalent to vice-chancellor, president or chancellor) of Imperial College, London, is a man of equable temperament, so that his protest last week that his Earth science department has been lumped, by the University Grants Committee (UGC), with the goats and not the sheep, carries extra weight. But what the quarrel (see page 489) signals is the emergence of a contradiction in the running of universities that has been on the cards for the past three years. After five years of confusion caused by the general shortage of funds, uncertainty has given way to an understanding that university resources for the prosecution of research would be concentrated on departments with international reputations.

Earth science departments are being tackled first (see *Nature* **332,** 101; 1988) only accidentally. Philosophy has already been rationalized, but the serious ructions will come next year with UGC's decisions about the research prospects of mainstream science departments latterly hard-pressed to fill student places. The essence of the contradiction is that the policy of concentrating university research resources on departments of high reputation is a reasonable way of apportioning inadequate resources within the existing British framework. That comes about because grant-making agencies reckon that universities will themselves cover the overhead costs of project research. But the policy is also, inevitably, one that limits the freedom of universities to mould themselves in their own lights. Ash's protest boils down to saying that an institution such as Imperial College must surely be incomplete without a substantial research programme in the Earth sciences. Who, in fairness, will say that he is asking for more than the head of a substantial science-based university is entitled to expect?

In reality, the only way in which clashes of this kind could have been avoided would, preferably, have been to have started from somewhere else or, alternatively, to have changed the framework in which British universities operate. For the past eight years, there have been more institutions than the funds available could adequately support. Allowing some of the weaker institutions to go to the wall would have been a solution, but in the British system would have required unpalatable political decisions. The merging of institutions, another solution, has proved uncommonly difficult (partly because of academic snobbery). But the pattern now in prospect of institutions which are lop-sided in ways they find intolerable is at once inequitable and unstable. The hope must be that, now that they can see the writing on the wall, they will defend themselves by seeking potential merger partners — and that the government will recognize that its schemes for centralizing control of the universities will saddle it with a host of uncomfortable problems in the years ahead. □

Academy and presidential panel issue AIDS reports

- Sweeping recommendations made
- Conflicting views on national commission

Washington

THE debate over the best way to address the mounting AIDS epidemic reopened last week with the release of reports by both the US Presidential Commission on the HIV Epidemic and a panel set up by the US National Academy of Sciences (NAS) and the Institute of Medicine. While the reports differ in many of their detailed recommendations, they both state that federal efforts to coordinate a response to the AIDS epidemic have been insufficient, slow and poorly led.

The NAS report is a follow-up to its 1986 document entitled Confronting AIDS, one of the first comprehensive studies to outline the catastrophic effects of the disease on US society and to make recommendations on research and health care. The revision does not repeat the earlier version's criticism that federal organization and the government's effort in education are "woefully inadequate", but does say that there is still an "absence of strong Federal leadership".

To fill the need for leadership, the NAS report calls for the creation of a national commission to replace the Presidential Commission on the HIV Epidemic, once that disbands after its final report to the White House later this month. The commission would report directly to the President for a renewable term of five years, and be active in directing efforts to combat AIDS. This is the second time the NAS has recommended the creation of a national commission.

Theodore Cooper, Chairman and Chief Executive Officer of Upjohn and chairman of the committee preparing the NAS report, says the committee rejected the idea of a single "AIDS czar", or of a separate agency to deal with AIDS, fearing that programmes already under way would be disrupted.

NAS's most urgent call is for a stronger approach to intravenous drug users, the population in which the virus causing AIDS is now spreading the most rapidly. The report points to the "gross inadequacy" of educational campaigns directed towards drug users, and the need for targeted educational programmes for black and Hispanic communities. The Academy condemns the inadequacy of drug treatment programmes and the dearth of facilities, and commends the evaluation of tactics such as handing out clean needles.

Other NAS proposals include the definition of HIV infection as a disease, in recognition that the stages of disease progression are a continuum; the establishment of a federal law to prohibit discrimination against people who test positive for infection; the foundation of a federal grant system to assist the states in financing health care for people with AIDS; and the removal of bans that prohibit the US Centers for Disease Control, as a government agency, from placing paid advertisements for AIDS education.

The NAS committee predicts that its 1986 estimate of $1,000 million a year for spending on AIDS education and health care will be too low by 1990, but recommends that proposals to spend more on research than $1,000 million should be balanced against other research priorities.

The draft report released last week by the Presidential Commission on the HIV Epidemic goes way beyond the recommendations contained in its preliminary report, released in February (see *Nature* **332**, 3; 1988). The 269 pages contain 579 recommendations based on the testimony of over 570 witnesses from government, the health-care industry, homosexual groups and academic institutions.

Contrary to predictions at its formation that the commission would toe the political line established by President Reagan, its report comes out against the mandatory testing of prisoners and immigrants, and suggests legislation to ban discrimination based on infection with HIV.

Like the NAS report, the commission's report calls for strong laws to protect the confidentiality of HIV test results, but recommends that the confidentiality rules be breached to notify the sexual partners of infected persons, health-care workers accidentally exposed to blood and state epidemiological authorities.

But the commission would also make it a criminal offence knowingly to transmit the virus, establish educational programmes in the schools at all ages and create a state funding pool to insure medically uninsurable people. The broad recommendations would cost more than $3,000 million a year extra for AIDS by 1990, according to the commission's chairman, retired Admiral James D. Watkins.

Reflecting the outspoken qualities attributed to Watkins over the past several months, the commission's report contains a final chapter written by Watkins himself which is likely to draw opposition from the more conservative commissioners, who will comment on the draft this week.

Watkins proposes that Congress should give the office of the US Surgeon General increased powers to act independently to avert a public health crisis once it has been identified by the President. The Surgeon General would construct plans and budgets to counter emergencies in consultation with pertinent agencies.

The high cost and drastic nature of the changes in the US management of the AIDS epidemic recommended by the NAS and the Presidential Commission inevitably mean that many of them will never happen. But Congress and the US Public Health Service (PHS) are both working to bring AIDS policy around.

A bill sponsored by Representative Henry Waxman (Democrat, California) which provides confidentiality and anti-discrimination protection to those infected will be considered in the House of Representatives this week. The PHS has just ended a week-long closed retreat involving its top AIDS officials to determine a future action plan. The officials compiled revised estimates of the incidence of AIDS, predicting 450,000 cases by 1993. Their plans for dealing with this staggering number will be released at the end of this month. **Carol Ezzell**

New research centres

THE National Science Foundation has announced a $2.2-million grant to establish one of three Biological Research Centers at Johns Hopkins University in Baltimore. The award, the largest of three, will support a new Institute for Biophysical Research on Macromolecular Assemblies to examine how assemblies of macromolecules — such as DNA, RNA and complex proteins — work together in the living cell.

To this end, the university has announced four major research thrusts: the regulatory interactions and structures in protein–DNA systems; protein stability and folding; molecular mechanics of the cytoskeleton, and structure and function of proton-translocating ATPases.

The University of California, Berkeley, will receive a $2-million, five-year grant to set up a centre for plant developmental biology. Director Michael Freeling says the money will be used to equip a cell biology laboratory, the centrepiece of which will be a state-of-the-art confocal microscope.

The facility will provide a "repository of cell biological techniques", says Freeling, with two full-time cell biologists to maintain the laboratory, develop new methods and teach techniques to researchers. Their assistance is needed, he said, as faculty members trained in molecular biology are being drawn by their research into the realm of cell biology.

The third centre will be at the University of Arizona in Tucson, for insect biology research. S.S. & M.B.

Voluntary HIV test launched for women

London

THE British government is to launch a programme of voluntary screening for the human immunodeficiency virus (HIV) in a sample of around 90,000 pregnant women over a one-year period. The number represents about a sixth of all pregnant women. The government may sanction involuntary testing, but not before widespread consultation on the legal and ethical issues, for which health minister Tony Newton has invited views.

By deciding to opt for voluntary testing, the government is accepting the recommendations of a working party set up last year under Dr Joe Smith, director of the Public Health Laboratory Service. Smith's report, *The Monitoring and Surveillance of HIV Infection and AIDS* (Department of Health and Social Security, London, £2.00) rules out involuntary, anonymous testing at first, "as we believe that the data required for the surveillance of HIV infection in the UK can be obtained by means of studies in which the consent of the patient has been secured".

The women would be given the choice to have the test done anonymously, in which case they would not be identified or learn the outcome. "In this way the bias due to individuals refusing to be tested may be reduced appreciably". Several of the country's leading epidemiologists have insisted that compulsory screening of all pregnant women is the only feasible means of obtaining an idea of the prevalence of HIV in the general population. It remains questionable, however, whether using a patient's blood for a purpose not specified is legal. The British Medical Association has said that such a move could amount to criminal assault.

For the study, three samples of pregnant women will be tested, each of the order of 20,000 to 30,000 women. Two samples will be from relatively high-risk areas — one in London and one in Scotland — and the third from a low risk area. The workshop group is confident that the size of the sample is adequate to obtain an accurate figure for the prevalence of the infection and that refusal rates are likely to be low because of the clinical value to the woman and the child.

Not everyone shares the group's confidence on the issue of refusal rates. Professor David Cox, of Imperial College, says that while the government's initiative was to be welcomed as a move in the right direction, the requirement for participants to volunteer leaves "an enormous question mark" over the eventual success of the study. **Simon Hadlington**

AIDS prevalence in pregnant women in France cause for worry

Paris

A RECENT study in of 274,647 pregnant women in France has revealed that as many as two births or abortions per day involve women carrying antibodies to the human immunodeficiency virus (HIV), the virus causing AIDS. Over one third of these women testing seropositive had no idea that they were at risk. The Health Ministry now has to decide what action, if any, should be taken in the light of these results, but is finding that the options are not simple.

The survey was carried out on the initiative of Professor Roger Henrion, an obstretrician at the Port Royal maternity hospital in Paris, with support from the Direction Générale de la Santé (public health directorate). Following a preliminary study in the Greater Paris area in 1987, Henrion extended his research to cover 239 clinics throughout France. The final sample represents over one third of the 778,000 births in France during 1987.

Although the health ministry has questioned the representativeness of the sample (because of geographical variation and because few private clinics were approached), the results give the best idea available of the spread of the AIDS virus among young French women. In particular, it confirms that the virus is so far concentrated in two metropolitan areas — Paris/Ile de France and Provence/Côte d'Azur — and that intravenous drug-users constitute the female population most at risk (66.7 per cent of the sample testing seropositive). Sexual contact is the next most common means of transmission.

An AIDS test is systematically proposed to pregnant women in only a minority of French clinics. (At present, only donors of blood, organs or sperm are required, by law, to be screened for HIV.) Other clinics propose an AIDS test only to patients identified as belonging to a high-risk group. The results therefore probably underestimate the true incidence of the AIDS virus among young women.

Although he thinks nationwide systematic screening is not yet warranted, Henrion believes there is justification for systematic voluntary screening where the risk of AIDS infection is high. Current statistics suggest that one out of two children born to seropositive women develop — and almost certainly die of — AIDS before the age of five. The speed with which a paediatrician can react with appropriate therapies when a baby falls ill is greatly increased if the child is known to be seropositive. This would be even more necessary should a vaccine or treatment for AIDS be found.

The health ministry wants to see Henrion's study extended to overcome statistical flaws. But the minister is not yet convinced that the cost of systematic voluntary screening of pregnant women is justified. **Peter Coles**

French AIDS figures highest in Europe

Paris

STATISTICS published by the World Health Organization (WHO) reveal that, with an estimated 63.5 AIDS sufferers per million inhabitants, France now has the highest incidence of known AIDS cases in the European Community (EC). The next highest is in Denmark (51.4 cases per million), followed by Belgium (33.9 cases per million).

Male homosexuals still comprise the majority of AIDS sufferers but the disease is spreading most rapidly among intravenous drug users. The new Health Minister, Claude Evin, decided last week to extend for a further year the sale of disposable hypodermic syringes without prescription. But a small survey of 300 addicts carried out for the Ministry of Health found that almost half of the sample still share syringes.

The French government Health Department (Direction Générale de la Santé) has found that, if current growth rates are projected, 21,101 cases of AIDS will have been diagnosed by the end of 1989, with as many as half needing hospital treatment.

At present almost 70 per cent of AIDS sufferers are treated in the five specialized clinics in Paris. A further six clinics have been set up in the provinces, but screening for the virus causing AIDS still remains a problem. Although anonymous free tests are available, with 100 centres to be opened in France by the end of 1988, the addresses of existing centres are not well publicized.

The cost of an ELISA test (about £14) is reimbursed only for those having valid state or private health insurance. Although the results of the test are not shown on the insurance form, the claim for reimbursement cannot be anonymous. This system is thought to dissuade some of those potentially at risk from taking an AIDS test. But Médecins du Monde, a charity run by doctors, has had to open a second Paris centre to cope with the demand for its no-nonsense, free, anonymous screening service. The state health service is now thinking of adopting a similar procedure in its new centres. Peter Coles

AIDS education could be working, but it is hard to tell

Berkeley

THERE is good news and bad news in an Office of Technology Assessment (OTA) report on the effectiveness of AIDS education, released yesterday*. The good news is an unprecedented change in sexual behaviour among homosexuals in San Francisco. The bad news is that education programmes aimed at homosexual men in other parts of the country, as well as efforts to reach intravenous drug users and young people, have been less successful. Moreover, education efforts have not been designed to make it easy to determine which elements are most effective.

The report cites abundant evidence that information alone does not persuade people to reduce their risk of contracting AIDS. In a Pittsburgh study, 91 per cent of homosexuals recognized anal intercourse as the most likely means of contracting human immunodeficiency virus (HIV), and believed that condoms would afford protection, but 65 per cent continued having anal intercourse, and nearly two thirds of those did not use condoms. Such failures are not limited to homosexual men. A San Francisco programme convinced 67 per cent of intravenous drug users that they should sterilize their needles with bleach, although only 15 per cent changed to safer sexual practices.

In contrast, San Francisco homosexuals have shown a dramatic reduction in risky sexual behaviour. From 1985 to 1987, the proportion of homosexuals engaging in receptive anal intercourse fell from 34 to 8 per cent. Half of the homosexuals in San Francisco are already infected with HIV, but the rate of new infections has dropped to less than 2 per cent per year.

Such dramatic risk reduction requires well-focused education programmes, as well as a shift in community norms, the report says. Programmes aimed at San Francisco homosexual men used extensive market research to find the best way to reach the desired audience, says Thomas Coates, of the University of California, San Francisco, who contributed to the report.

The programmes not only provided information, but also instructed people in behaviour-changing skills, involved community members in the education process and attempted to develop an environment supportive of behaviour change.

But even such apparently successful programmes have not been rigorously evaluated. The report notes that the success of a programme "may be more a function of the characteristics of the participants than of the programme itself". This is apt for San Francisco, where the homosexual community is well established and is made up primarily of middle-class professionals who may be more likely to be receptive to AIDS education.

According to OTA, little is known about the attitudes and sexual practices of inner-city black homosexuals, and some studies suggest they are less well-informed about AIDS. One Detroit survey found that only 13 per cent of black homosexuals knew that AIDS is transmitted through blood and semen.

OTA recommends carefully aimed educational programmes combining mass-media and face-to-face instruction. But for such a programme to work, the report says its messages must take into account "the language, literacy level, and cultural sensitivities of the people to whom they are targeted". To design an effective educational effort, OTA advises researchers first to acquire a better understanding of the behaviour and attitudes of all AIDS risk groups. The report recommends that the Centers for Disease Control, in collaboration with state governments, should conduct a systematic variation of media messages, followed by an evaluation of their relative effectiveness.

Marcia Barinaga

**How Effective is AIDS Education?* (Office of Technology Assessment, Washington, DC, June 1988).

Erratum

Jacques Beckers has indeed left the US National New Technology Telescope programme (*Nature* **333**, 6; 1988) but not to direct the European Southern Observatory's Very Large Telescope (VLT) project, which is being run by Daniel Enard and Massimo Tarenghi. Beckers will work for ESO on imaging techniques, especially interferometry, which will be an essential element of the VLT. □

Saving Japanese rocks out at sea

Tokyo

A SMALL armada of ships armed with helicopters and launches has begun a $240-million operation to salvage two small but vital Japanese rocks poking a few tens of centimetres out of the western Pacific. But all the effort may be in vain.

At high tide the rocks are all that remain above sea level of the coral island of Okinotorishima, Japan's southernmost territory 1,700 km south of Tokyo. If they are washed away, Japan will lose about 400,000 square kilometres of its 200-mile exclusive economic zone and with it all rights to fishing and minerals in the area. So an attempt is being made to save the 'island' by building protective walls around the rocks.

But even if the rocks survive, Japan may not be able to claim the 200-mile zone around them. Soji Yamamoto, professor of international law, points out that article 121 of the yet-to-be-ratified United Nations Law of the Sea Convention states that "rocks which cannot sustain human habitation or economic life of their own shall have no exclusive economic zone".

Yamamoto suggests that, if Japan wants to press a claim, a lighthouse, wireless station and port should be installed on the island before the convention comes into effect (36 developing nations have signed the convention but 60 must sign for ratification and developed nations are at present holding out). "Territory is not something that exists naturally", Yamamoto says, "it must be obtained by a country's efforts".

David Swinbanks

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Ships surround Okinotorishima while a helicopter lowers an iron tripod onto a wave-breaking wall of tripods being built around one of the island's two remaining rocks. After the two little 'islands' have been surrounded with walls 50 m in diameter, the enclosed areas will be partially infilled with concrete to reinforce the walls. But care must be taken not to pave over the rocks with concrete, because under international law they would then be disqualified as parts of a natural island. The rocks, both of which are eroded at the base, must stand on their own (Asahi Shimbun Photo Service).

Human Frontiers Science Program on table at Toronto summit

Tokyo

JAPAN's Human Frontiers Science Program will be on the table at the Toronto summit of Western nations to be held later this month. The summit is expected to focus on economic issues and the decision to include Frontiers, made at a pre-summit meeting in Toronto on 5 and 6 June, is largely symbolic. But it may help the Japanese government ministries and agencies involved to realize their dream of turning the programme into a major multilateral research organization.

Frontiers was first proposed by the Ministry of International Trade and Industry (MITI) in 1986 as a huge international bioscience project into which Japan would pour thousands of millions of dollars to develop new technology for the twenty-first century. But the Science and Technology Agency has since assumed prime responsibility for the programme, emphasis has switched to pure science and the proposed scale of the budget has been considerably reduced. Nevertheless, agency officials still entertain great ambitions for the programme.

Japan will, if necessary, put up all the funds for Frontiers in the initial phase, according to Tateo Arimoto, director for planning in the agency's Science and Technology Policy Bureau. The agency and MITI hope to get about $20 million for Frontiers in fiscal year 1989 and these funds will probably be fed into a central organization to distribute international grants and fellowships, Arimoto says. But as the grants will be for three years and the fellowships for two, the Frontiers budget will have to rise two to threefold by 1991.

Other Japanese ministries may join the programme from fiscal year 1990 and discussions are now under way between MITI, the agency and the Ministry of Education, Science and Culture. But Arimoto is also optimistic that other nations will join.

He says West Germany, France, Italy and the European Community have expressed willingness to contribute. Britain, on the other hand, is not expected to give any funds but is willing to provide scientists and access to her research laboratories. Representatives of the US National Science Foundation, the National Institutes of Health and the National Academy of Sciences have also said they will cooperate. But William Graham, head of the White House Office of Science and Technology, has made it clear that the United States will not contribute funds in the initial phase, although he also welcomed the programme. And Arimoto says that whereas in the past representatives of other countries were sceptical and said "you will do it", the attitude is now changing to "we will do it".

Arimoto foresees a loose international organization in the initial phase, perhaps with governments contributing funds to central Frontiers organizations in each participating country. But by the early 1990s, he hopes that a multilateral organization similar to the European Molecular Biology Organisation will be established. And this will help the flow of funds across Japan's borders, he says.

Japan's Prime Minister Noboru Takeshita is not expected to discuss any of these details at the summit. But inclusion of Frontiers is "very important politically", according to Arimoto, and will strengthen the agency's hand in negotiating the Frontiers budget with the Ministry of Finance later this year.

David Swinbanks

Cuts in tax concessions for Australian R&D

Sydney

CUTS in tax concessions for investment in research and development (R&D) have been announced in the Australian federal government's May economic statement. The 150 per cent tax concession approved in 1985 will continue for the previously agreed period to 1991. The concession will then drop to 100 per cent from June 1991.

Professor Frank Larkins, president of the Federation of Australian Scientific and Technological Societies (FASTS) believes that the federal government's decision may result in a reduction in industrial R&D unless other initiatives are set in motion.

Although there has been only a gradual increase in R&D investment since the concession were first introduced, it has mainly benefited companies with sizeable in-house research rather than smaller companies or universities. Helen Trincker, spokesperson for the Australian Vice-Chancellors Committee, says: "In the 1985–86 period, of the proportion of money invested under the 150 per cent tax cut scheme, only 4 per cent went to universities while 96 per cent of companies invested in their own institutions. To counteract this, we are creating a Council for Business/Higher Education Cooperative which will, hopefully, forge links between industry and universities. The trouble with industry is that they must recognize that research can be slow and they may not get returns on their investments for a couple of years."

Professor Larkins believes that greater incentives must be offered, now that tax concessions have been reduced, so that R&D investment does not fall. "In the May statement, company tax was dropped from 49 to 39 per cent. This will mean an increase in profits, part of which could finance research. All companies with more than five employees should be required to contribute at least 1 per cent of their turnover directly to R&D spending."

Tania Ewing

Scandinavian killer algae outbreak

London

A SERIOUS outbreak of poisonous algae in the seas around southern Scandinavia has stopped short of a widespread ecological disaster. Three weeks after the problem began, strong winds and currents have contributed to its recession, diluting the concentration of viable algae by 75 per cent in two days.

The poisonous gelatinous algae *Chrysochromulina polylepsis* have devastated marine life in the Kattegat and Skagerrak seas and cost the Norwegian fishing industry some £120 million, including the loss of 500 tonnes of salmon. Disaster committees have been set up by the Swedish, Danish and Norwegian governments and extra funds injected into marine research laboratories to accelerate the control of the algae, which was doubling overnight and spreading in enormous clumps at 25 km a day.

As a result of the episode, the Norwegians are now considering setting up many more land-based fish farms, where the environment can be better controlled. The Swedes are urgently looking into the development of ecological early-warning systems, having recently been the victims of Chernobyl, of acid rain from other industrialized countries, and now of a possible water-borne pollution which could have caused the algae to flourish.

A spring bloom of marine algae is a natural event resulting from changing conditions in sea water. Particular algae establish themselves by exerting inhibitory effects on rival species as a result of their secretions, a potent toxin in the case of *C. polylepsis*. The surviving species are those best adapted to the prevailing conditions.

Proliferation is encouraged by increased concentrations of nitrate and phosphate washed into the sea as agricultural fertilizers and sewage. Rare strains of algae can become dominant as a result of water pollution by noxious chemicals, simply by growing more rapidly. Pollution carried by rivers from eastern Germany is being blamed for the provision of this perfect *C. polylepsis* medium, but this is probably an oversimplification.

Rosalind Cotter

Applied science to lose out in UK university subject reviews

London

A BITTER dispute has broken out between two of Britain's most respected senior academics over the outcome of the recently completed review of university Earth sciences departments, with concern being expressed that departments with a bias towards applied science will lose out in future subject reviews. The rector of Imperial College, London, Professor Eric Ash, says that in the case of his college's geology department the criteria adopted by the review committee were "largely irrelevant" and that by recommending that the department cut its student numbers by 80, to 175, and shed 10 out of the 35 academic staff, the committee "got it dreadfully wrong".

Sir Peter Swinnerton-Dyer, chairman of the University Grants Committee, which commissioned the review, defends the review committee's decision. In a letter to Ash he says that the department is perceived as badly managed, that its staff engage in too much consultancy work and that it underachieves in research performance. Furthermore, Swinnerton-Dyer says that publication and citation analyses for the department, the country's largest in Earth sciences, "tend to confirm the national (review) committee's view of the department as complacent and living on past glories".

Responding to this, Ash points out that several influential applied journals are not quoted in the citation index, nor are proceedings of symposia. On the staff's activity in consultancy work, Ash says that the average number of consultancy days a member of staff does in a year is 10, with 70 the maximum number for a single member. "Applied geologists are useful to the geology-based industry. Consultancy is financially advantageous to the staff; it is also academically advantageous to the staff and our research students".

Ash says that the case of his college's geology department highlights deficiencies in the review process that will have serious implications for future reviews of university science, distorting the balance between pure and applied science.

Reviews of physics and chemistry are under way, with biology to begin soon.

When the Earth sciences review was completed in March (see *Nature* **332**, 101; 1988), three categories of department were announced: type M (for 'mainstream'), type I (for 'interdisciplinary') and type J (for 'joint honours teaching only' and without expensive facilities). Departments of types M and I were further classified into groups 1 or 2, with the former having provision for greater student numbers. There was general surprise that Imperial's geology department found itself in type I, group 2.

Ash is convinced that by its positioning in group 2, the department is being penalized for its emphasis on applied research. To support his case, he presents a detailed analysis of performance indicators that he feels are relevant to an applied department. Of the 33 departments reviewed, Imperial scores consistently highly in measures such as earnings per member of staff in research grants and contracts, number of full-time equivalent students, student-to-staff ratio and number of postgraduate students. Ash concludes that the review committee lacked a fundamental appreciation of the working of an applied department. Swinnerton Dyer, however, remains unswayed.

Simon Hadlington

Nuclear agreement

Tokyo

THE new and controversial nuclear agreement between the United States and Japan was finally approved by the upper house of the Diet on May 25, and is expected to take effect in the autumn.

The agreement gives Japan approval for 30 years to transport nuclear fuel, most of which comes from the United States, across international frontiers and to reprocess it.

Originally, Japan had planned to return plutonium from European reprocessing plants by air freight, using Anchorage, Alaska, as a refuelling point, which provoked objections in the US Congress. But plans have now been revised and the air transports are expected to fly non-stop to Japan, probably over the North Pole.

Japan intends to transport 25 tonnes of plutonium over a period of 20 years starting in the 1990s, which will require flights every month. David Swinbanks

Nuclear arms race at sea

Washington

To coincide with the Reagan/Gorbachev summit and the ratification of the INF treaty by the US Congress, pressure groups in the United States are trying to focus attention on the nuclear arms race at sea. Thus a report* recently released claims to be the first complete and publicly available inventory of the world's naval nuclear forces.

Nuclear weapons at sea — some 16,000 of them worldwide — make up nearly one third of the world's total nuclear stockpile, according to the report's authors, William Arkin and Joshua Handler, defence analysts at the Institute for Policy Studies in Washington, DC. The authors also contend that increases in the world's seafaring nuclear arsenals constitute a "hidden arms race".

Arkin and Handler are concerned that, while the INF agreement would eliminate land-based intermediate-range nuclear missiles, nuclear arsenals at sea have often been neglected in arms control even as they have become increasingly important in overall nuclear strategy.

The report, which describes the number, type and location of nuclear weapons and nuclear-capable ships and aircraft in the navies of the United States, the Soviet Union, Britain, France and China, explains that nearly half of the world's nuclear naval arsenals are not covered by any arms control negotiations. Among these "uncontrolled" armaments are approximately 3,300 anti-submarine warfare (ASW) nuclear weapons.

Introducing his report, Arkin argued that discussion of the nuclear arms race at sea has been hampered by the lack of accessible authoritative information and that governments have further clouded the issue by secrecy, "neither confirming nor denying" reports of weapons on ships.

The report is especially concerned with the prospects for sea-launched cruise missiles (SLCMs), which it says will be used to replace the intermediate-range missiles to be eliminated by the INF treaty. SLCMs are already a stumbling block in the START negotiations on strategic weapons at Geneva. Describing them as the "most dangerous weapons in production", Arkin says SLCMs are potentially important in regional conflicts.

According to the report, the US and Soviet Union have nearly 550 nuclear SLCMs deployed on 75 surface ships and 94 submarines, with plans to deploy hundreds more on at least twice as many platforms at sea. **Seth Shulman**

* "Nuclear Warships and Naval Nuclear Weapons: A Complete Inventory", Neptune Paper No. 2, May 1988, published jointly by the Institute for Policy Studies and Greenpeace.

US–Soviet space cooperation

Washington

FOR cooperation in space as for arms control, the Moscow summit meeting last week between President Ronald Reagan and General Secretary Mikhail Gorbachev yielded modest, positive results, but no real breakthroughs. As expected, the meeting's joint communique included an agreement to expand civil space cooperation between the two countries. But the US delegation did not take up Gorbachev's explicit proposal (announced to the US press before the summit) of a joint Mars mission.

The most tangible item in the summit agreements will be the "exchange of flight opportunities". Scientific instruments from each country will fly on the other nation's space missions. The communique does not name specific collaborations, but several are already tentatively scheduled. In the first of such arrangements, NASA (National Aeronautics and Space Administration) plans to send a US-built ozone mapping instrument on the 1990 launch of a Soviet Meteor weather satellite. The instrument would be a slightly modified version of the Total Ozone Mapping Spectrometer (TOMS) that has been returning data about the deterioration of the ozone layer from the US Nimbus 7 spacecraft.

The Soviets, in turn, plan to include a receiver aboard the US Mars Observer spacecraft to be launched in 1992. The Soviet instrument would relay images to Earth from a Soviet Mars balloon flight in 1994. Also on the docket are plans to launch an X-ray telescope developed by the United States and Denmark on a Soviet satellite in the mid-1990s. Some dozen or more similar exchanges are also reported to be under consideration.

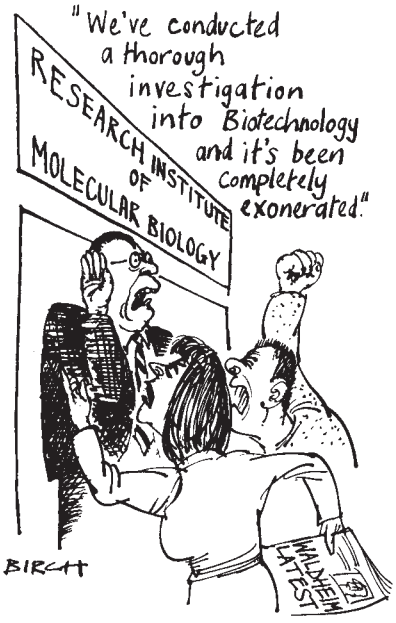
The Soviet and US delegations also committed themselves to expanded exchanges of space data and scientists, but there was no movement on a joint Mars mission. Gorbachev had formally proposed a joint venture only days before the summit, a prospect which, according to some sources, sent the White House into a panic. The current US administration has consistently opposed the idea, and the Pentagon has expressed concern that such a collaboration would allow the Soviets to gain access to sensitive US technology.

A State Department spokesperson in Washington acknowledged that the summit resulted in only "a modest expansion" of previous goals on cooperative space ventures outlined in an April 1987 agreement between the two countries. Officials at NASA and the State Department stressed that it did, however, represent a small step forward. **Seth Shulman**

Out of the woods for Viennese molecular biology?

Vienna

ABOUT fifty student demonstrators were on hand to mark the official opening at the end of May of a laboratory that is set to act as a focus for modern biological research in Vienna. The protesters, mainly students of Vienna University opposed to "gene and reproduction technology", were particularly exercised by the fact that the new Research Institute of Molecular Pathology (IMP) is a joint venture of two commercial concerns. But by 1991, the



university hopes to have several institutes on the same site, creating a 'Biozentrum'.

Genentech, the Californian biotechnology company with aspirations to be a pharmaceutical company, and Boehringer Ingelheim, the West German pharmaceutical giant with which Genentech has had increasing links over the past few years, own and finance IMP, although the city of Vienna has contributed 20 per cent of the 700 million schillings (£35 million) for the first five years.

Although the ultimate purpose of IMP is to produce product ideas for the two companies in the area of cancer therapy, the institute's brief is to carry out basic research. The companies believe IMP is the first basic research institute to be set up as a joint industrial venture.

The institute, with a total of 80 staff, is directed by Professor Max Birnstiel and has a scientific advisory board that includes Dr David Baltimore, director of the Whitehead Institute in Cambridge, Massachusetts. Five senior scientists have now been recruited within Europe and will all be in post by the end of the year.

Among the attractions for new staff are high salaries, generous transferable pensions and the provision of a service department for protein and DNA sequencing and synthesis. Senior appointments are for two years plus three years rolling tenure, a system Birnstiel would like to see spread in Europe. While IMP has an initial 12-year life-span, it can be considered as "permanent", says Clifford Orent, Genentech's director of international operations. Permanency, however, is likely to depend on the successful emergence of product ideas. Birnstiel hopes that these will emerge largely because his staff are conscious of the need and because of personal contacts with company scientists that he will strongly encourage.

Despite the demonstration by students, whose concerns were of the general kind that has emerged with some force in neighbouring West Germany, the activities of IMP seem unlikely to be affected. The students' concerns, however, may yet have some political impact.

While Professor Hans Tuppy, rector of Vienna University until last year when he became Austria's Minister of Science and Research, was one of the speakers at the institute's opening, other senior politicians were noticeable by their absence. Plausible excuses were available, but the suspicion remained in some quarters that genetic manipulation may be set to become a public issue in Austria and that some politicians wanted to keep their distance for that reason. **Peter Newmark**

Californian centre for excellence named

Berkeley

A JOINT venture between Stanford University and the University of California's Berkeley and Santa Barbara campuses has been selected as one of five 'centers for excellence' to be funded by Sematech, the semiconductor manufacturing research consortium formed to try to restore US leadership in semiconductor manufacturing technology.

The other centres named are the University of Arizona for contamination and defect control, a group of New Jersey universities for plasma etching, the University of New Mexico for metrology and the Massachusetts Microelectronics Center for single-wafer processing. Each centre will receive between $500,000 and $1.5 million per year from Sematech for five years.

Sematech says the purpose of the centres is "to provide a focus for development of manufacturing programs within graduate schools of US universities".

Marcia Barinaga

Postal ministry offers hope for German computer network users

Munich

THE future has brightened considerably for the users of academic computer networks in West Germany. After months of anxiety about the prohibitive and unpredictable cost of leased telephone lines, the West German postal ministry will announce this week that it will provide an all-purpose network of the X.25 type for academic use at fixed cost.

When IBM stopped paying for the European Academic Research Network (EARN) at the end of 1987, a scramble began across Europe to maintain the level of service EARN had provided (see *Nature* **328**, 752–753; 1987). National organizations took over from IBM and lines were reshuffled to adapt to the wide variations in telephone rates. West Germany lost three of its five international lines, not only because of its high rates, but also because customers were charged according to the volume of their traffic.

Although university researchers in West Germany found other channels for their domestic correspondence, such as using the facilities of the *Grossforschungseinrichtungen* (Large Research Establishments or AGF), they faced an uncertain future from an apparently unsympathetic postal ministry.

But some time in the past year, postal ministry attitudes changed. Postal officials apparently stopped seeing the research community as just another special interest group with unreasonable expectations, and came to regard researchers as users who could help to demonstrate the ministry's new products and services that might also interest the more lucrative business market. Deregulation and the beginning of competition in telecommunications probably helped to bring about the change.

The postal ministry is reported to be offering researchers a network that will initially carry data at a speed of 64,000 bits per second (b.p.s), with the opportunity to increase the speed to 2 million b.p.s. The network will be available from mid-1989 and will be administered by the German research network (*Deutsche Forschungsnetz* or DFN). The tariffs for the new network have yet to be determined but DFN is expecting a significant reduction from current charges.

The postal ministry has also asked DFN to develop technical standards for a pilot broad-band network capable of carrying data at 140 million b.p.s. Such networks depend on fibre-optics cables currently being installed by the postal ministry all over West Germany. They can be used, for example, for remote communication with supercomputers or for extremely detailed image transmission and manipulation. There are already a number of experimental broad-band networks in existence in West Germany, and apparently an increasing number of scientists are applying to the postal ministry for permission to install such facilities.

The future for international scientific networks in Europe remains uncertain despite the progress in West Germany (see *Nature* **329**, 775–778; 1987). The postal ministries (PTTs) of the various European countries tend to see themselves as autonomous bodies with little interest in cooperation. EARN director Dennis Jennings of University College, Dublin, while greeting the "dramatic change" in West Germany as "fantastic", is concerned about "national boundaries that serve as international tariff walls". One attempt to lower the walls is called Managed Data Network Services (MDNS). It is being developed jointly by the PTTs to offer users 'one-stop shopping', that is, the opportunity to use international lines while dealing only with the local PTT. The first MDNS node is currently being installed in Holland, with others expected to follow soon.

Jennings views MDNS with caution, warning that it "does not [yet] exist" and will have a difficult birth, judging by EARN's experiences in negotiating with the various PTTs. Meanwhile, EARN continues to negotiate for corporate support to improve its facilities. The next year will bring further developments in this rapidly changing field. **Steven Dickman**

A long and rocky road to reversing rainforest destruction

Washington

POOR, short-sighted economic policies are to blame for the high rate of deforestation in developing countries, according to a report* by the World Resources Institute (WRI), a Washington-based think-tank concerned with the influence of public policy on the environment. The decimation of up to one-third of the developing world's forests since 1950 is only partly attributable to the practice of shifting cultivation, the conversion of forest land for agriculture and fuelwood gathering by local populations, the report concludes.

The report points a finger at the well-meaning but mistaken policies adopted by developing countries to deal with their problems of poverty, urban overcrowding and trade imbalances.

Tropical countries in the developing world often take for granted the long-term benefits their forest resources can offer, and sell logging rights too cheaply. Logging concessions are usually not granted on a competitive basis, and are structured to turn over too frequently for the forests to regenerate, WRI finds. To compound the problem, internal political instability and corruption often mean that revenue received from logging concessions does not go directly into federal coffers.

Tax and other economic incentives designed to spur development by stimulating the forest industry can backfire. In some countries, the government may actually be paying concessionaires to come and take trees, because the amount of money that the government receives by selling access to its forests is negated by the loss in tax revenue and the administrative costs of managing the contracts, according to the WRI study. To keep some of the profits at home, many countries have encouraged the development of wood-processing industries.

But the inefficiency of the plants and overly generous tax breaks and loan agreements mean that the processed products often bring in less money than would the export of the raw logs, while increasing the amount of forest that is destroyed.

To counteract the acceleration in the rate of tropical deforestation caused by misguided public policies, WRI recommends reforms in the granting of logging rights and the restriction of economic incentives. WRI is calling on the Food and Agriculture Organization (FAO) of the United Nations to incorporate these goals into the Tropical Forestry Action Plan developed by the FAO, the World Bank and the United Nations Development Programme last year. **Carol Ezzell**

■ The Smithsonian Institution is also turning its attention to rainforests. The Smithsonian has cooperated with the World Wildlife Fund to create "Tropical Rainforests: A Disappearing Treasure", an exhibition intended to familiarize the public with the present beauty and the tragic future of these forests. The exhibition opened on 21 May in the S. Dillon Ripley Center in Washington, DC.

The exhibition is a warning in high-technology guise. It is intended to alert the public to the links between deforestation and the adaptation of land for agriculture, cattle ranching, timber operations and development projects. It also chronicles the effect of rainforest destruction on the world debt crisis and the growing human overpopulation. **Lisa L. Lyles**

* *The Forest for the Trees? Government Policies and the Misuse of Forest Resources*, World Resources Institute, Washington, DC, 1988.

Science and creationism

SIR—*Nature*'s even-handed analysis of the US Supreme Court decision on the Louisiana Balanced Treatment Act (*Nature* **327**, 643: 1987) has, as would be expected, brought responses from the viewpoints of creationists (Reginald T. Chelvam, *Nature* **331**, 10: 1988) and naturalists (Andrew W. P. Roberts, *Nature* **331**, 476: 1988). Under the sometimes emotional and personal reactions of these groups, there are two very different ways of looking at reality, and the genesis of this conflict lies in neither side acknowledging, much less seriously examining, the other's fundamental precepts in their own right.

From a naturalistic viewpoint, in which the totality of reality is to be found in the physical realm, the human activity of science is the way to truth, because there is no other reality beyond the physical and because science is the most successful means to study the physical universe. The attempted inclusion, therefore, of a deity or other supernatural activity in a naturalistic worldview is firmly and logically denied. That is, the naturalistic philosophy of scientism, which recognizes only that accessible to sciences as real, by definition denies anything beyond the physical. Science itself is neutral with respect to the supernatural.

From a theistic viewpoint, by which I mean a viewpoint which accepts a literally 'super-natural' deity as a valid aspect of reality, a central question arises from the nature of the interaction between the spiritual and physical realms. While there are a great variety of such worldviews, fundamental Christianity, from which a 'scientific creationist' viewpoint arises, is one that perceives a reality comprised of both natural and supernatural realms. In this perspective, both the physical and spiritual are legitimate aspects of reality and hence a naturalistic viewpoint that denies the spiritual is unacceptable.

Unfortunately, most of the interaction between proponents of these two worldviews is not at the level of these basic premises but rather several steps removed. The usual result is a shouting match rather than a constructive examination of each other's fundamental premises which would then provide a basis upon which to discuss the consequences of holding to these different worldviews. It is ironic, but perhaps not unexpected, that some theists (scientific creationists) appropriate the great credence of science in an attempt to justify and validate their position by calling on the name and methods of science rather than on the name and authority of their God. The naturalist's denial of any role of faith in their understanding of reality is likewise a curious matter as any worldview ultimately rests, in faith, on unproven presuppositions. We thus have the proponents of naturalistic and theistic worldviews inappropriately claiming authority over territory belonging to the other, while denying the other's viewpoint as valid except through their own interpretations.

This is a sorry state indeed for those involved in the fray. Could we not now move, in a manner science so often claims for itself, to a more objective appraisal of such difficult issues as origins, starting from a respectful analysis of the basic presuppositions of each party rather than from emotional states removed by several orders from fundamental precepts?

ANDREW P. WHIPPLE
Taylor University,
Upland, Indiana 46989, USA

Academic freedom

SIR—I have followed the controversy, particularly in relation to academic freedom, about the Education Reform Bill championed by the UK Secretary of State for Education and Science, Mr Kenneth Baker (*Nature* **332**, 386; 1988).

Guaranteeing academic freedom and sanctioning job security or permanent tenure are two separate issues. In this debate the two have unfortunately been mixed up. Academic freedom is vital for universities and must be preserved at all costs. What is academic freedom? It is the freedom granted to the academic community (students and teachers) to pursue knowledge as they see best without being subject to external pressure from church, government or economic forces. The choice of courses of study, the problem of research in various branches of learning, the procedure of evaluation and the award of degrees ought to be the monopoly of the university community with no outside interference. With such a sense of academic liberty, the academic community will in turn generate a sense of discipline and duty towards society at large. Within the ambit of academic freedom lie the means to preserve the memory of hard-fought battles for truth.

A teacher should be as free to give his or her opinion on controversial issues as any other citizen of a free society. To hold a certain opinion is a right granted to an individual as a member of a free society. So long as an opinion is not detrimental to academic growth, the holder is not liable to lose his job. Whether an individual's opinion or action is an impediment to the furtherance of academic growth should be judged by the academic community itself. This is true academic liberty. Any reform in education, however well intended, that does not guarantee to safeguard academic freedom makes the academic community subservient to the ignorance of politicians and to bureaucratic arrogance. In the words of Cardinal Newman: "A university is a place of concourse, whither students come from every quarter for every kind of knowledge. It is a place where inquiry is pushed forward and discoveries verified and perfected and rashness rendered innocuous and error exposed by the collision of mind with mind, and knowledge with knowledge." That is possible only with unabridged academic freedom.

A.N. MALVIYA
Centre de Neurochimie,
5, rue Blaise Pascal,
67084 Strasbourg, France

Green beliefs

SIR—Your leading article "Greens against genes" (*Nature* **332**, 667; 1988) cannot pass unchallenged. The implication that political environmentalists are "anti-intellectual" is both unjustified and unjust. Far from being anti-intellectual, the environmentalist party in Britain, to which the German Green Party in large part owes its intellectual heritage, is unique in its perception of the consequences of science and technology, an issue to which the other political parties pay lip service but are unable or unwilling to do justice.

The traditional distinction between science and technology must on this point, as elsewhere, be recognized and given due weight. Research into genetics is a science, genetic engineering is a technology. The consequences of science can only be to extend the sum of human knowledge, the consequences of a technology for the human race as a whole may be good or evil.

Combustion, to take an example, has been studied by chemists and we know much about it. Some of the technology resulting from this knowledge has been beneficial, such as the development of fire extinguishers that do not use chlorofluorocarbons; some has been disastrous, such as the dirty and excessive burning of fossil fuels, which must either be reduced or will ultimately cease under potentially catastrophic circumstances. Similarly, genetic engineering may have beneficial consequences (a cure for AIDS is much to be desired) but it may also have evil consequences if it allows politically controlled or financially motivated technologists to engender monstrosities.

The green parties of Europe favour the beneficial uses of technology and condemn those that are harmful. Far from being anti-intellectual, this stance is both clearly thought out and socially responsible, true to the humanitarian spirit of the great scientists of the past on whose work our present understanding rests.

RICHARD J. BIRD
50 Highbury, Jesmond,
Newcastle upon Tyne NE2 3EA, UK

Why the pressure to publish?

Despite technological change and the increasing difficulty of managing the scientific literature, there seems to have been little change as yet in the pattern of publication. But that could change quickly.

MOST people will agree that there is far too much of the scientific literature, but nobody seems to have a convincing remedy. Indeed, the general opinion is that the numbers of journals, and the physical bulk of the successful among them, will continue to grow more or less as they have done for the past several decades. But there is just a chance that present conventions will change under the influence of previously unknown forces, among which new techniques for processing information are only one.

That technology will have an important influence is not open to dispute, although its influence is easily exaggerated and the timing of change remains guesswork. The benefits, while still largely unrealized, are easily imagined. When it becomes commonplace for manuscripts submitted for publication to be handled electronically until the very act of putting ink on paper is reached (and even that may not be necessary), the process of publication could be much faster than at present. But that is only the most obvious feature of what may be in store. When galley proofs are eliminated, or at least replaced by electronic signals which are checked by electronic comparison of the original and the edited data-streams, much aggravation will be avoided. So too, with the help of laser printers and an electronic version of the published paper, will be the annoyance of waiting for reprints to arrive. But none of this will, in itself, diminish people's wish to be published.

Nor will the developing technology eliminate the printed page. Both readers of, and contributors to, journals have a common interest in the survival of ink on paper. For readers, the interest is that of assimilating a variety of information quickly; people's reading habits differ, but there must be something in the view that the most useful information is gathered with the corners of the eyes, while the sight of a printed article is much more effective than the title in telling whether it can be safely left unread.

Contributors mostly welcome printed versions of their articles both because they also have an interest in seeing that what they write should be assimilated and for the natural (and forgivable) warm glow that these tangible proofs of effort bring. So, on the face of things, the volume of the published literature could continue indefinitely to grow a little more quickly than the number of active researchers.

But the forces in the other direction, far from negligible even now, may be growing even more quickly. One is the large but uncounted cost of the printed literature. It can easily be seen how the world could be spending far in excess of $1,000 million a year on publications; it is simply a matter of believing that there are 5,000 journals selling 1,000 subscriptions at a cost of $200, modest by many standards.

If the cost (including page charges, where appropriate) of producing and supplying the world's basic research journals were greater than, say, the budget of the National Science Foundation in the United States, nobody would be surprised, but a great many people (not all of them administrators) would be asking whether all that money might not be diverted to more practical uses. And that arithmetic cannot take account of the largely unsung efforts of referees and the honorary editorial boards whose work keeps the journals going.

The problems of the management of the literature exert similar pressures. No doubt the great abstracting journals such as *Chemical Abstracts* would be unusable were they not already available in a form that can be searched by machine, unsatisfactory though the product may often be. For librarians, the challenge is not so much that the bulk of the literature is growing, but that the complexity of, say, finding an item, or telling which items are to be found, increases still faster.

Even when the search for a particular article has succeeded, there remains the difficulty of telling what it says. It is not so much that science has been balkanized by specialization, but that the conventions of the trade impede communication even with would-be readers. Mostly, the conventions are well-meant, but their consequences can be disastrous. People whose results disagree with others already published do not say so openly, no doubt from fear of seeming trouble-makers. Descriptions of experimental procedures are copied in full from laboratory notebooks, but without saying how they differ from well-known precedents (which would often be sufficient), perhaps from fear of being hauled before some investigating committee to face allegations of incomplete disclosure. And there is an increasing and distressing tendency towards po-faced reticence about the motives and meaning of even the most interesting work. It is as if people fear that if they say directly what they think the significance of their work may be, they will be blackballed by all the promotion boards and prize-award committees before which it is their ambition to appear.

Problems of comprehension are likely, in the years ahead, to force even the most severe journals into different habits. Already there are essentially academic journals that have taken to publishing reviews of current topics, some of which are as snappy as if they were journalism. The result is to make the literature a little more accessible. Only the most staid will be able indefinitely to resist this trend. At the same time, there is bound to emerge a clearer distinction between material of topical and less topical interest.

But, the realists will say, none of this will seriously dent the impetus that chiefly sustains the publication process — the determination of the appointments and promotions boards with which the academic system is littered to judge people primarily by their publications. But is that necessarily the case?

Publications have understandably been the chief grist for these mills — they are tangible products of people's work, they can be counted objectively and assessed with at least the appearance of objectivity. But there is also a growing sense of frustration with published output as a measure of a person's standing. It is not so much that a bibliography may give a false impression of what a person has accomplished but that, with the laconic conventions of the trade now practised, it may also underrate him or her.

Yet there seems little prospect that moderate changes in the present system will be quickly accepted: a modest proposal, at Harvard, that candidates for appointments should be judged on the basis of a handful of published works only has apparently led to the cry "We want to see the poor papers too!" Those who distrust too much reliance on bibliographies should welcome this reaction; it can only hasten the time when people are judged by their colleagues by what is known of them personally. But since the realists are probably correct in their opinion that the eagerness to publish, not to mention the use of the Minimum Publishable Unit as the quantum of what is made public, is sustained by the activities of the boards, radical change there could quickly bring about radical change in the pattern of publication.

John Maddox

Molecular evolution

Origins of the AIDS virus

David Penny

THE RNA viruses are our best examples of evolution in 'real' time. Because of their high error rate during replication[1] they show, as expected[2], both a high diversity in their host and a rate of evolution about a million times faster than DNA-based organisms. The human immunodeficiency virus (the AIDS virus) fits this category, and two new papers extend our knowledge of the evolution of retroviruses. On page 573 of this issue[3], Smith *et al.* give a detailed analysis of relationships for strains of human immunodeficiency viruses (HIV-1 and HIV-2) and the simian immunodeficiency virus (SIV). And in a recent issue of *Proceedings of the National Academy of Sciences*[4], McClure *et al.* consider a wider group of retroviruses but use a simpler analysis.

Smith *et al.*[3] use nine amino-acid sequences of the viral protein from isolates of HIV-1, HIV-2 and SIV. The cornerstone of their work is an alignment of these nine sequences using the positions of conserved cysteine sites, non-conservative amino-acid changes, and then nucleotide replacements. Though not used as a criterion, potential sites of glycosylation were conserved. This alignment is used to reconstruct a tree which is shown in its unrooted form in Fig. 1.

Most authors, even today, still stop at this point and 'believe' the first tree they find. But Smith *et al.* tested the tree by forming many subsamples and constructed a new tree for each sample. The samples always lead to the same tree, giving confidence that the tree would not change as more data are collected — that is, convergence to a single tree had occurred. In addition, the authors used a second tree-building method which gives the same tree. It is possible to quibble that random subsets, as in bootstrapping and jackknifing, may be better[5,6] but in the present case the result appears robust.

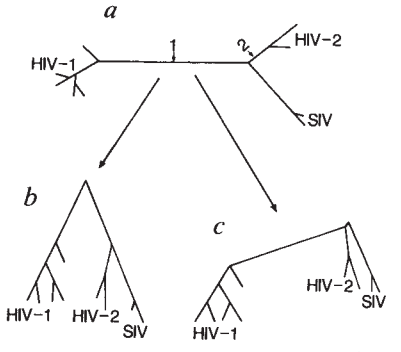


Fig. 1 *a* The unrooted tree of Smith *et al.*[3]. This unrooted tree has been tested by finding the same tree from different subsets of sequences. Smith *et al.* place the root near the centre (marked 1) to equalize rates of evolution. This leads to the rooted tree in *b*. If HIV-1 is a much more virulent strain of HIV-2, it may well have a much higher rate of evolution, and the root may therefore still be near the point marked 2, giving the rooted tree *c*.

Where should the root of the tree be placed? (The subsampling tested only the unrooted tree.) The authors' choice for the root, in their words the "minimum evolutionary assumption", is to place the root so that the average rate of nucleotide change was constant — the molecular-clock assumption. With the root here, the simplest assumption is the unexpected conclusion that monkeys have been infected by a human virus, not humans infected by a monkey virus. (In addition, this leaves open the question of the source of the three viruses.) Alternatively, HIV-1 and -2 could have arisen independently from SIV (see last week's News and Views article by Carel Mulder[7]).

The assumption of transmission from humans to monkeys cannot be rejected out of hand, but Smith *et al.*[3], although favouring equal rates, point out that other positions for the root are possible. Many would argue that other positions are better.

Many authors seem to assume that the rate of molecular evolution is a DNA-given constant. With retroviruses, it is expected that most of the accumulated changes will occur in the RNA phase. RNA synthesis, or DNA synthesis from RNA, is expected to have a high error rate of the order of one in 10^3 or 10^4 bases[1], in contrast to the rate of about one in 10^9 for DNA synthesis from DNA.

If, for example, the root was at position 2 in Fig. 1, then transmission may have gone from monkey (SIV) to human, first as HIV-2. It is possible that HIV-1 is derived from HIV-2 as a particularly virulent strain with a corresponding short incubation period, and consequently a fast rate of evolution. (Each cycle from DNA to RNA and back to DNA would have a high chance of mutation.) This is like the old argument that the number of generations per unit time will affect the rate of molecular evolution. We need to know much more detail about the life-history of the virus, or be able to determine the root independently.

A common method for determining the root is to use an 'outgroup' (sequences that are more distantly related to SIV and HIV) and finding where they join the tree. Some of the new SIV strains[7] seem to be good candidates for an outgroup. Alternatively, McClure *et al.*[4] used a wide variety of retroviruses, some of which could be used as an outgroup. However, the main interest of these authors is in the relationships of mammalian and avian retroviruses. They used nine full viral sequences and another eight partial sequences in their study.

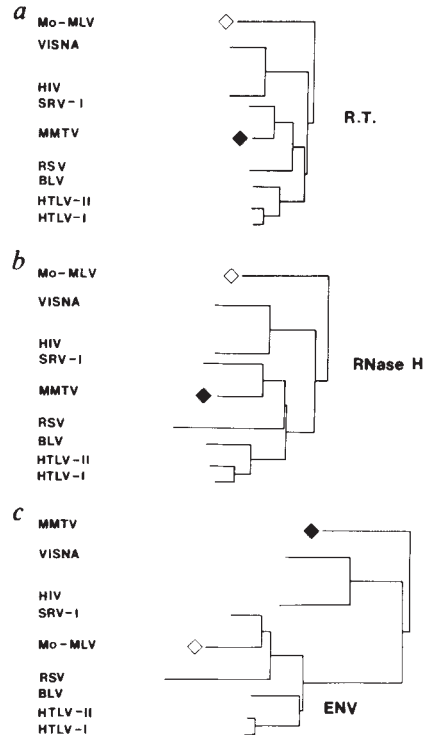


Fig. 2 The tree of McClure *et al.* *a*, Reverse transcriptase (R.T.) tree; *b*, RNAse H tree; *c*, envelope-protein (ENV) tree. MMTV, mouse mammary tumour virus; Mo-MLV, Moloney mouse leukaemia virus; VISNA, visna lentivirus of sheep; HTLV, human T-cell leukaemia virus. (From ref. 4, with permission of Dr R.F. Doolittle.)

The results of McClure *et al.*[4] show that enzyme sequences generally change more slowly, and give more similar trees, than non-enzymatic sequences. The envelope proteins, which are evolving faster, give a different tree (Fig. 2). The authors attribute the different trees to recombinational events. There is a problem in that the authors give no test for robustness, so it is not yet possible to evaluate quantitatively the evidence for such a conclusion.

Do the different trees result because sequences are too short to give convergence (sampling error)? The lengths of some sequences are quite short and in our experience[6] we would be surprised if the faster-evolving envelope-protein sequences give the same tree as the slower. What happens when the faster-evolving sequences are fitted to the reverse-transcriptase tree (Fig. 2*a*)? If the fit to this tree is almost as good as to Fig. 2*c* then the differences may just be random effects from building trees with short sequences. If the fit is markedly worse, then recombination could be a reasonable explanation. Further, McClure *et al.* use only one

method to reconstruct trees, so the difference may be an artefact of that method. The authors do not discuss the location of the root of the tree.

MacClure *et al.* built a composite tree from 4 sequences for 17 viruses[4]. This allows them to identify five groups of retroviruses. Some groups have been recognized previously (C- and E-type leukaemia viruses) where others are more diverse. Again the results are interesting, although no quantitative analysis is possible.

Comparing the papers of Smith *et al.*[3] and McClure *et al.*[4] highlights the advantage of a thoroughly tested tree as a basis for analysis. And, as Smith *et al.* remark, "The reconstruction of the phylogenetic history of HIV-1 and HIV-2 is an important component for the understanding of the epidemiology of the spreading disease, AIDS." □

1. Reanney, D.C. *Nature* **307**, 318–319 (1984).
2. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983).
3. Smith T.F. *et. al. Nature* **333**, 573–575 (1988).
4. McClure, M.A. *et al. Proc. natn. Acad. Sci. U.S.A.* **85**, 2469–2473 (1988).
5. Felsenstein, J. *Evolution* **39**, 783–781 (1985).
6. Penny, D. & Hendy, M.D. *Molec. biol. Evol.* **3**, 403–417 (1986).
7. Mulder, C. *Nature* **333**, 396 (1988).

David Penny is in the Department of Botany and Zoology, Massey University, Palmerston North, New Zealand.

Astrophysics

The shifting centre

Virginia Trimble

EUROPEAN and American astrophysicists have been in the habit of supposing that their Japanese colleagues are thinking and doing more or less the same things they are, but perhaps a year or so later. At a recent conference* it was made abundantly clear that this supposition is badly out of date and that the centre of astronomical research is no longer unambiguously in the mid-Atlantic. A major part of the varied programme consisted of new, wholly or partly Japanese results in the areas of neutrino astronomy, X- and γ-rays, the microwave background and data-processing technology.

The solar neutrino problem[1] is almost as old as many of the conference participants. From 1968–69 onwards, the unique ^{37}Cl experiment operated by Raymond J. Davis Jr and co-workers has persisted in detecting only about one-third as many neutrinos with energies above 0.814 MeV as the standard solar models predict. Most of us ceased to doubt the essential correctness of the experiment many years ago.

Confirmation is nevertheless very welcome. The Kamiokande II experiment, though designed primarily for other purposes, records the energy and incident direction of neutrinos above 9 MeV and so can detect a portion of those emitted by ^{8}B decay in the Sun. Yoji Totsuka (University of Tokyo) reported that the current count rate from the direction of the Sun is about 0.1 event per day, whereas the standard solar model predicts 0.3. Allowing for the still-limited statistics, this translates into ruling out a flux larger than 62 per cent of the standard model at the 90-per-cent confidence level and ruling out zero flux at the 95-per-cent level. John Bahcall (Institute for Advanced Study, Princeton), discussing solar neutrino theory immediately thereafter, began by saying he had waited 20 years to hear Dr Totsuka's talk.

The Ginga X-ray satellite[2] has been returning data in the 4–30 keV band since summer 1987. Among its products are the first X-ray spectra above 2 keV for a number of quasars and Seyfert galaxies (active galactic nuclei or AGNs). Hajime Inoue and Katsuji Koyama (Institute of Space and Astronautical Science, Tokyo) described both the continua and the line spectra. The continua are generally well fitted by power laws whose slopes we very much want to know to decide whether we can account for the nearly-isotropic X-ray background by adding up the faint, distant AGNs, or whether we must invoke hot intergalactic gas as well. The few AGNs studied earlier seemed to have a canonical slope near 1.7, too soft (meaning too few of the highest-energy X-rays) to match the 1.4 slope of the background. Ginga now finds a range of slopes, 1.5–1.9 for quasars, 2–3 for BL Lac objects and 1.4–2.4 for Seyfert galaxies. That is, a few spectra are hard enough to match to the background radiation, but most are not. Several of the objects are variable on timescales from minutes to days.

These X-ray continua are thought to arise from non-thermal processes like Compton or synchrotron radiation. But many of the detected AGNs also showed an iron emission line between 6 and 7 keV, revealing the presence of hot gas. The exact energy of the (de-redshifted) line tells us how many electrons still orbit the atoms. In the quasars 2C273 and E1821+643, the iron must be fairly heavily ionized and thus close to the central quasar engine. The Seyfert lines, on the other hand, appear at a slightly lower energy, implying X-ray fluorescence by neutral gas. The case of NGC1068, a Seyfert 2 (the less spectacular kind), is particularly interesting. Its X-ray spectrum greatly resembles that of the nearby stellar source Vela X-1 during eclipses, when we see X-rays from the central source indirectly, reprocessed by surrounding cooler gas. Thus the Ginga data support a recent hypothesis that Seyfert 2 galaxies are much like other more active objects, but with our view of the central engine obscured, presumably by an accretion disk.

Ginga also recorded X-rays from supernova 1987A in energy bands of 6–16 and 16–28 keV beginning in July[3]. The hard X-rays, also seen by the Kvant module on the Soviet Mir satellite[4], are thought to be scattering-degraded γ-rays emitted as ^{56}Co decays to excited states of ^{56}Fe. But the turn-on was 6 months or more earlier than expected. Even more unanticipated were the apparent variability of these X-rays and the simultaneous turn-on and flaring in the softer band.

Yasuo Tanaka (Institute of Space and Astronautical Science) and Kuniaki Masai (Nagoya University) provided updates on the data and their interpretation. Tanaka suggested that the observed X-rays can be deconvolved into two components. One has the spectrum expected from degraded γ-rays obscured by 10^{25} atoms cm^{-2} and has remained nearly constant from July to February. (This is, in itself, surprising, and requires the decaying ^{56}Co to have been mixed through most of the supernova ejecta rather than being concentrated in a layer near the bottom.) The other component has a thermal spectrum and has undergone a couple of minor flares and a major one in January when it brightened by a factor of 6 to 10^{38} erg s^{-1} and heated up from 12 to 50–60 keV.

The January flare included a 6.7–6.9 keV iron line, indicating at least some hot-gas emission. It was obscured by less than 10^{23} atoms cm^{-2} and so must have arisen well outside the zone of the hard source. These points present difficulties for a model in which the X-rays come from a central pulsar or pulsar-driven nebula (Franco Pacini, Astrophysical Observatory, Florence). But the obvious alternative of ejecta colliding with pre-existing dense gas is not fully satisfactory either. The peak flux declined in only a couple of days, requiring very rapid cooling of the shocked gas and thus expansion at the unlikely speed of 10^{9} cm s^{-1}. The absence of associated radio emission and still softer X-rays is also puzzling. Masai noted the collision with a single dense cloud rather than a symmetric shell reduces the discrepancies.

Photons at much higher energies (greater than about 1 TeV) from SN1987A have also been widely predicted (ref. 5 and many others) on the belief that high-energy cosmic rays must be accelerated early in supernova events. Just possibly these TeV γ-rays have been seen,

* Twentieth Yamada Conference: *Big Bang, Active Galactic Nuclei and Supernovae*, 28 March–1 April 1988, University of Tokyo.

according to Tadashi Kifune (University of Tokyo), reporting for a Japanese–Australian–New Zealand collaboration. Their Cerenkov light detector saw a modest increase above background in the direction of the supernova during the January X-ray flare. If real, it corresponds to a luminosity of about 3×10^{38} erg s^{-1} above 3 TeV, persisting for only a couple of days.

The search for spatial and/or spectral lumpiness of the relict 3-K microwave background radiation is also about 20 years old. The most persuasive spectral data so far come from a February 1987 rocket flight mounted jointly by Nagoya University and University of California, Berkeley[6]. Analyses of the data have continued since initial theoretical discussions[7] found that the excess 480- and 710-μm radiation could be attributed either to inverse Compton scattering of longer wavelength photons or to thermal emission from dust heated by pregalactic starlight.

Toshio Matsumoto (Nagoya University) showed the observed fluxes as a function of ecliptic latitude. For the 100–200 μm range, there is a correlation with interstellar hydrogen density and with zodiacal emission (interplanetary dust), as seen by the Infrared Astronomy Satellite. At wavelengths longer than 400 μm, the fluxes are essentially isotropic, with possible fluctuations on a scale of degrees. The residual excess emission, according to Matsumoto, is then best fitted by dust emission at $3.7 \times (1+z)$ to $4.4 \times (1+z)$ K, where z is the redshift at which the dust was heated, and the exact temperature depends on how the dust absorption coefficient varies with wavelength. The necessary energy input is sizable in any model, amounting to about 10 per cent of the $(1+z)^4$ eV cm^{-3} contained in the total microwave background.

Finally, several of the contributed papers and posters touched upon issues of modern technology and data-processing techniques. For instance, Tsuneaki Daishido and colleagues from Waseda University, Nobeyama Radio Astronomy Observatory and Sony Corporation's Information Systems Research Center reported on a technique for rapid processing of radio signals from arrays of antennae.

Their prototype processor, intended to look for transient radio sources, is capable of 2.7 billion (10^9) operations per second and has been used for studying solar radio bursts. I ought to be able to tell you more about it, as the authors provided a seven-page, 1988 reprint, complete with extensive circuit diagrams. As I examined the pages, Satio Hayakawa (Nagoya University) remarked that Americans have been complaining that Japanese researchers do not send state-of-the-art information promptly. He suggested the alternative interpretation that they are sending, but we are not receiving very well. This must be true, for I cannot tell you even the name of the journal that published this forward-looking paper — in Japanese. □

1. Bahcall, J.N. & Ulrich R.K., *Rev. Mod. Phys.* **60**, 297–372 (1988).
2. Makino F. *et al. Astrophys. Lett. Commun.* **25**, 223 (1988).
3. Makino F. *et al.* IAU Circular No. 4532.
4. Sunyaev, R. *et al. Nature* **330**, 227–229 (1987).
5. Gaisser, T.K., Stanev, T. & Halzen, F. *Nature* **332**, 314 (1988).
6. Matsumoto, T. *et al. Astrophys. J.* (in the press).
7. Hawakawa, S. *et al. Publ. astr. Soc. Japan* **39**, 941 (1987).

Virginia Trimble is professor of physics at the University of California, Irvine, California 92717 and visiting professor of astronomy at the University of Maryland, College Park, Maryland 20742, USA.

Quantum optics

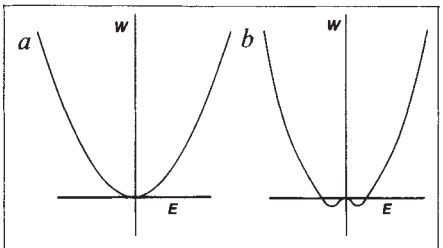
Noise-driven phase for lasers

Peter Knight

PHASE transitions between different output states are common for lasers. The most important is that between ordinary fluorescent behaviour of the laser medium and coherent amplifying laser action. Another occurs at the onset of chaotic fluctuations in the output field at higher powers. J.H. Hannay and D. Field suggest elsewhere in this issue (*Nature* **333**, 540–542; 1988) that a third transition can be driven by noise fluctuations of the output electric field. They liken this transition to the Curie–Weiss transition in a ferromagnetic material, in which cooperative interactions between individual atomic magnetic dipoles makes the whole ensemble very sensitive to small thermal fluctuations.

The laser is often regarded as a nonlinear oscillator driven by an external pump mechanism which attempts to maximize the excitation of the atoms forming the source of radiation. The excited atoms form a fluctuating polarization which in turn creates the laser electric field. If the pump rate is insufficient to overcome the relaxation of the atoms and field, the electric field dies away. The decaying transient fields from each atom are randomly phased and add to form mere noise with a zero mean amplitude. When the pump rate is raised sufficiently to overcome the losses in the laser cavity, gain is possible and a steady-state electric-field amplitude can be maintained. The critical pump rate here defines a first threshold, which is that necessary to provide sufficient excitation to balance the losses.

A second threshold occurs (Haken, H. *Phys. Lett.* **A53**, 77; 1975) when, for very high levels of pumping, the evolution of the laser electric field becomes completely chaotic, an effect seen in many nonlinear systems. The phase transition discussed by Hannay and Field in this issue is distinguished from both of these through its reliance on fluctations in the atomic polarization of the laser. Their results indicate that small changes in the laser parameters can lead to very large changes in the average intensity of the laser output.



The first phase transition of a laser. Below threshold (*a*), the 'potential' (*W*) of the laser is minimum at mean electric field (*E*) zero; corresponding to chaotic, incoherent fluorescence. Above threshold (*b*), the well bifurcates, giving two minima, one of which is picked spontaneously to generate a non-zero laser electric field, responsible for the coherent laser output. In the new phase described by Hannay and Field, the form (width) of the bifurcated well is extremely sensitive to polarization fluctuations.

The method used by Hannay and Field is a straightforward extension of the theory used by Haken and H. Risken (see Haken, H. *Laser Theory* 2nd edn, Springer, Berlin, 1984) to describe the transition from noise to order in passing through the first threshold. In this standard approach, 'Langevin' equations, which model the fluctuations of the polarization, inversion (the degree of excitation of the atoms) and the laser electric field about their mean values, are used to define a 'Fokker–Planck' equation which describes how the probability distribution of the electric-field strength evolves. This distribution is governed by the size of the electric-field loss and gain in the laser. As a function of electric field, the probability distribution behaves rather like a potential well which can have a single minimum (*a* in the figure) when the laser is below threshold, or a double minimum when it is above threshold.

The evolution of the laser electric field behaves rather like a ball rolling in this potential and below threshold will roll to the bottom of the well. Above the threshold, the well bifurcates (*b* in the figure): two minima symmetrically displaced from zero represent the positive and negative stable equilibria of the ball. Spontaneous symmetry breaking allows

the ball into one rather than the other minimum.

This symmetry breaking occurs in the laser. Increasing the pump parameter until the first threshold is exceeded changes the laser system so that the most probable electric-field strength changes from zero at the bottom of the single-well potential (a random fluorescent output) to a non-zero stable solution at the bottom of one of the two wells, corresponding to a steady-state laser output.

By making a minor alteration to the Langevin equation that describes the polarization fluctuations, representing these as δ-correlated (with zero memory) white noise, Hannay and Field show that the double-well-potential characteristic of laser action is extraordinarily sensitive to the scale of atomic fluctuations. They demonstrate this sensitivity both through numerical simulations of the fluctuating dynamics and by analytic approximations which exploit the simplifications achieved when the fluctuations in the polarization dominate all others.

It is to be expected that fluctuations would allow the electric-field evolution to sense both wells simultaneously by exploring the entire range of the double-minimum structure within a strong fluctuation. In the new work, however, the situation is made more complicated because the mean field and the shape of the potential well are also dependent upon the scale of the fluctuations.

The obvious parallel between this laser phase transition and the better known Curie–Weiss phase transition in ferromagnetism is pointed out by Hannay and Field. Both are described by a probability distribution which depends on its own mean or mean square which leads to its having a discontinuous response to a continuously varying pump or external magnetic field.

The effective 'potential' in which the laser electric field moves is expected to be sensitive to fluctuations. What is surprising about this new result is how very sensitive the evolution turns out to be when only a minor variation in atomic polarization is assumed. It suggests that the output of a laser system is to a large degree dependent on polarization fluctuations, and as these can be altered by the environment, for example by collisions, it should be straightforward to study this experimentally. Clearly, it is necessary to see whether these predictions survive in more realistic versions of the model, which would include both the appropriate decay rates in the evolution equations and improved treatment of the fluctuations, not treating this merely as additive white noise. □

Peter Knight is in the Blackett Laboratory, Imperial College of Science and Technology, London SW7 2BZ, UK.

Foamy retroviruses

A virus in search of a disease

Robin A. Weiss

RETROVIRUSES are broadly grouped into three subfamilies: RNA tumour viruses or oncoviruses (themselves subdivided into B-, C- and D-type viruses); slow viruses (lentiviruses); and foamy viruses (spumaviruses)[1]. Oncoviruses long held the lion's share of the interest. But then, the causative agent of AIDS, human immunodeficiency virus (HIV), was classified as a lentivirus, which turned the spotlight onto slow, neuropathic lentiviruses such as visna virus. Spumaviruses remained the Cinderella group, having no important diseases associated with them. Now, Rolf Flugel and his group at the German Cancer Research Centre in Heidelberg have completed their analysis of the genome of human foamy virus (HFV), which has interesting properties[2–4].

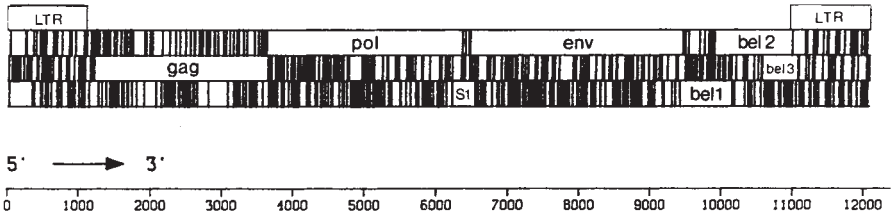
The genome organization of the HFV provirus is shown in the figure. It is just over 1,200 base pairs long and, like all retroviruses, has *gag, pol* and *env* genes in 5′ to 3′ order. The *pol* gene overlaps the *gag* gene by 22 nucleotides and seems to be expressed by a translational frameshift, in the same way as Rous sarcoma virus. The *gag* gene is longer than most retroviral *gag* genes and can include the proteinase sequence at its carboxy terminus. The 5′ half of the HFV genome containing *gag* and *pol* is slightly more similar to mammalian C-type oncoviruses than to other retroviruses, particularly in the reverse transcriptase. It is noteworthy, too, that foamy virus reverse transcriptase activity prefers Mn^{2+} to Mg^{2+} for its divalent cation[5], again resembling mammalian C-type viruses.

The 3′ half of the HFV genome, on the other hand, is more remininscent of lentiviruses than of C-type oncoviruses. There are three open reading frames (ORFs) between *env* and the 3′ long terminal repeat, which Flugel *et al.*[2] call *bel*-1, -2 and -3. But they leave us bemused about the derivation of *bel*. There is also a possible ORF, S1, between *pol* and *env*. It is not yet known whether the *bel* genes encode regulatory proteins, but the amino-acid sequences of *bel*-3 and the 3′ ORF protein of HIV-2 have regions of similarity[3]. If a complete infectious provirus can be constructed, the functions of these genes could be ascertained by site-specific mutagenesis.

Foamy viruses were so named because in cell culture they induce multinucleated giant cells with numerous vacuoles, giving the syncytia a foamy appearance. The vacuoles are derived from rough endoplasmic reticulum and possibly Golgi, and contain budding and mature virions. Foamy viruses were first described more than 30 years ago from macaque kidney cell cultures[6]; there are about nine known distinct serotypes of simian foamy viruses[7]. Many were isolated as latent infections in simian and chimpanzee brains[8] in the course of research on the disease Kuru. Foamy viruses have also been described in cats, cattle and hamsters. They were classed with the RNA tumour viruses and visna virus in 1971 when reverse transcriptase was detected in the virions[5], four years before the term retrovirus was coined[9]. Flugel *et al.* call their virus human spumaretrovirus (HSRV)[2], but I think the simpler term human foamy virus (HFV) is better and conforms with the established usage of simian foamy virus (SFV).

HFV was the first human retrovirus to be described. The first isolate was made in cultures of nasopharyngeal carcinoma cells[10,11], and further putative human isolates were made from various tissues[12–15]. Because HFV is serologically related to SFV-6 of chimpanzees and no antibodies specific to HFV were found in a sample of 256 human sera of diverse provenance, Gajdusek's group[16,17] queried whether HFV is a naturally occurring human virus. But Epstein, Loh and associates[18,19] subsequently detected about 5 per cent seropositivity for HFV in East Africa and Pacific island human populations. As chimpanzees are phylogenetically closer to man than to monkeys, it seems plausible

1. Weiss, R.A., Teich, N.M., Varmus, H.E. & Coffin, J. *RNA Tumour Viruses.* (Cold Spring Harbor Laboratory, New York, 1985).
2. Flugel, R.M., Rethwihn, A., Maurer, B. & Darai, G. *EMBO J.* **6**, 2077–2084 (1987).
3. Maurer, B. & Flugel, R.M. *FEBS Lett.* **222**, 286–288 (1987).
4. Maurer, B., Bannert, H., Darai, G. & Flugel, R.M. *J. Virol.* **62**, 1590–1597 (1988).
5. Parks, W.P., Todaro, G.J., Scolnick, E.M. & Aaronson, S.A. *Nature* **229**, 258–260 (1971).
6. Rustigian, R., Johnston, P. & Reihart, H. *Proc. Soc. exp. Biol. Med.* **88**, 98–116 (1955).
7. Hooks, J.J. & Gibbs, C.J. *Bact. Rev.* **39**, 169–185 (1975).
8. Gajdusek, D.C., Rogers, N.G., Basright, M., Gibbs, C.J. & Alpers, M. *Ann. N.Y. Acad. Sci.* **162**, 529–550 (1969).
9. Baltimore, D. *Cold Spring Harb. Symp. quant. Biol.* **39**, 1187–1200 (1975).
10. Achong, B.G., Mansell, P.W.A., Epstein, M.A. & Clifford, P. *J. gen. Virol.* **40**, 175–181 (1971).
11. Epstein, M.A., Achong, B.C. & Ball, G. *J. natn. Cancer Inst.* **53**, 681–688 (1974).
12. Young, D., Samuels, J. & Clarke, J.K. *Archs Gesamte Virusforsch.* **42**, 228–234 (1973).
13. Stanček, D., Stančekova, M., Janotka, M., Hnilica, P. & Oravec, D. *Med. Microbiol. Immun.* **161**, 133–144 (1975).
14. Cameron, K.R., Birchall, S.M. & Moses, M.A. *Lancet* **ii**, 796 (1978).
15. Werner, J. & Gelderblom, H. *Lancet* **ii**, 258–259 (1979).
16. Brown, P., Nemo, G. & Gajdusek, D.C. *Infect. Dis.* **137**, 421–427 (1978).
17. Nemo, G., Brown, P.W., Gibbs, C.J. & Gajdusek, D.C. *Infect. Immun.* **20**, 69–72 (1978).
18. Muller, H.K. *et al. J. gen. Virol.* **47**, 399–406 (1980).
19. Loh, P.C., Matsuura, F. & Mizumoto, C. *Intervirology* **13**, 87–90 (1980).
20. Hooks, J.J. & Detrick-Hooks, B. *J. gen. Virol.* **44**, 383–390 (1979).



Genomic organization of human foamy virus. Vertical bars, stop codons in the three reading frames. Scale in base pairs. (From ref. 4, courtesy of R. Flugel.)

that *Homo sapiens* and *Pan troglodytes* harbour closely related foamy viruses. Now that gene probes are available, it is opportune to compare the simian and human viruses molecularly, to devise more specific serological tests, and to examine human infection more closely.

There is scant evidence that foamy viruses are pathogenic in man or animals. SFV has caused transient immune suppression in rabbits[20]. Perhaps most interesting has been the repeated isolation of foamy viruses from a cluster of patients in Slovakia suffering from de Quervain subacute thyroiditis[13]. The agent was not at the time recognized as a foamy virus, although this later became apparent[15]. Even if the relationship between thyroiditis and HFV holds up (and there have been no subsequent reports), caution must be used in exploring whether the presence of the virus is cause or effect (activation of latency). The claim[13] that patients seroconverted to produce neutralizing antibodies to their unidentified virus suggests that an aetiological role for HFV is worth pursuing. □

Robin A. Weiss is at the Institute of Cancer Research, Chester Beatty Laboratories, Fulham Road, London SW3 6JB, UK.

Weather systems

Complex or just complicated?

Itamar Procaccia

THE dynamical behaviour of the wind in the atmosphere is much simpler than anything that could be anticipated, according to Tsonis and Elsner on page 545 of this issue[1]. They conclude that one can accurately model the wind dynamics using just eight coupled ordinary differential equations. Such a set of equations can be integrated on a home computer. This conclusion is contrary to common wisdom; one expects the dynamics of the atmosphere to be governed by the field equations of fluid mechanics (partial differential equations) whose solution defies the biggest supercomputers available. Tsonis and Elsner's suggestion, if correct, would lead to a conceptual breakthrough of fundamental importance, changing the way that we predict the weather. Their conclusions are in agreement with previous suggestions of two other groups[2,3]. Nevertheless, the evidence is not clear-cut, and further tests must be conducted to substantiate the conclusions.

Certain systems, with dynamics governed by partial differential equations, exhibit complexity that is no worse than that of a few coupled nonlinear ordinary differential equations. These systems have been studied extensively as part of the effort to understand the onset of chaos in dynamical systems (see the News and Views article by S. Wiggins in last week's issue[4]). In dynamics, chaos means that the solutions of dynamical systems depend extremely sensitively upon the initial conditions; nearby trajectories diverge exponentially, and thus long-range predictability is lost. The fact that complex behaviour of this type (termed deterministically complex) is possible in 'simple' systems of nonlinear ordinary differential equations is important. It means that dynamical complexity does not necessarily stem from the involvement of many dynamical variables. (Of course, complex behaviour can appear in a highly complicated system with many degrees of freedom.) It thus becomes essential to understand why some systems exhibit deterministically complex behaviour and when they are just complicated.

In principle, any fluid requires an infinite number of variables to describe its dynamics. A fluid has space- and time-dependent velocity, temperature and pressure fields. These fields can be expanded in a complete set of functions, and then the amplitudes of these functions obey an infinite set of ordinary differential equations[6]. It can happen, not too far from equilibrium, that a fluid is in a stable stationary state, meaning that small perturbations about that state eventually decay to zero. It can also happen that upon displacing the fluid further from equilibrium, only a few of these modes become unstable and develop interesting nonlinear dynamics on 'slow' macroscopic timescales (seconds or minutes). All the other modes are dynamically irrelevant and are enslaved by the few relevant modes. This is the regime in which a few ordinary differential equations can accurately describe the dynamics, although the dynamics are chaotic.

It has been found experimentally[7] and understood theoretically[8,9] that to observe this regime, the system's physical size should be small, and the departure from equilibrium, as measured by, say, the Reynolds number, should be small as well. As the physical extent of the system increases, the separation of timescales between 'slow' relevant modes and 'fast', irrelevant modes disappears. Also, a large system cannot respond coherently; defects and grain boundaries become significant, and the possibility of description with a few ordinary differential equations is lost. The number of relevant modes increases rapidly with the Reynolds number and at atmospheric conditions there should be nothing left of the low-dimensional chaotic behaviour.

Yet Tsonis and Elsner[1] describe in this issue an analysis indicating that even the atmosphere has only about eight important modes that govern its dynamics. This can stem from one of two reasons. Either there is something fundamental in atmospheric behaviour that we do not understand, which leads to an unexpected simplification of its dynamical behaviour; or Tsonis and Elsner's method of analysis does not apply or is outside its range of validity. The technique of Tsonis and Elsner was devised by Grassberger and me to estimate the dimension of 'attractors' of deterministically complex systems[10].

The idea is that in such systems, the orbit is confined to a subset of phase space, called a strange attractor. This attractor has a dimension D that is usually fractal. Moreover, this object can be embedded in an abstract space of dimension $d>D$. (Rigorously, d has to be at least $2D+1$.) There is a way to manipulate the data such that it yields the coordinates of points on the object in the d-dimensional embedding space. By counting how many pairs of points $C(l)$ there are whose distance is smaller or equal to l, one finds that this number scales as l_D. (Think about a line

100 years ago

We are glad to see that the Government Bill for the Promotion of Technical Instruction is down for second reading as first order of the day on June 14. It is an enabling Bill, giving powers to localities, if they think fit, to apply local rates to the purpose of promoting technical instruction.

In Clause 6, "technical instruction" is defined to mean "instruction in the principles of science and art applicable to industries, and in the application of special branches of science and art to specific industries or employments." This definintion appears good, so far as it goes, but in our opinion it does not go far enough, for it does not specifically include the commercial subjects and modern languages.

From *Nature* **38**, 121; 7 June 1888.

embedded in ordinary space and points sprayed on the line. Here, $C(l) \sim l$ and the generalization to higher dimensions is obvious.) A plot of $\log C(l)$ against $\log l$ should yield a straight line of slope D. Random data always fill the embedding space, be it of any dimension d. This method is extremely useful for analysing experimental data exhibiting low-dimensional chaos. Unfortunately, when D increases, a successful application calls for an exponentially increasing amount of data. Roughly speaking, for a statistically meaningful scaling law $C(l) \sim l^D$ over a decade of lengths l, about 10^D points in the data set are needed. For $D \approx 7$, about 10^7 data points are needed. Tsonis and Elsner were able to collect only 3,960 points in their wind measurements. It is thus possible that the amount of available data is simply insufficient to provide meaningful calculations of a dimension.

Studies in natural conditions often yield an amount of data that cannot be increased. How can the dimension algorithms be supplemented to distinguish deterministic from random data? One approach, suggested recently, is based on the idea that deterministic strange attractors are dense with unstable periodic orbits. The chaotic motion itself can be conceived of as a random walk between these periodic orbits. The system follows a path close to an orbit for a while, until its instability causes the system to follow a path close to another orbit. My colleagues and I have used algorithms to extract these periodic orbits straight from experimental data[11]. In discrete time, the number of points N_n belonging to orbits of period n should increase exponentially with n. Random data do not have such orbits; moreover, the amount of data needed to extract (at least low n) periodic orbits is not large. Finally, the knowledge of the available periodic orbits and their stability can yield information about the dimension of the attractor as well[12]. The demonstration of the existence of periodic orbits in data of the type analysed by Tsonis and Elsner would be an important piece of evidence in favour of its originating in deterministic complexity rather than being just complicated. □

1. Tsonis, A.A. & Elsner, J.B. *Nature* **333**, 545–547 (1988).
2. Fraedrich, K. *J. atmos. Sci.* **43**, 419 (1986).
3. Essex, C., Lookman, T. & Nerenberg, M.A.H. *Nature* **326**, 64–66 (1987).
4. Wiggins, S. *Nature* **333**, 395–396 (1988).
5. Ciliberto, S. & Gollub, J.P. *J. Fluid Mech.* **158**, 381 (1986).
6. Maron, E. & Procaccia, I. *Phys. Rev.* A**34**, 3221 (1986).
7. Ahlers, G. & Behringer, R.P. *Phys. Rev. Lett.* **40**, 712 (1978).
8. Coullet, P., Fauve, S. & Tirapegui, E. *J. Phys., Paris* **46**, L787 (1985).
9. Chate, H. & Manneville, P. *Phys. Rev. Lett.* **58**, 112–xxx (1987).
10. Grassberger, P. & Procaccia, I. *Phys. Rev. Lett.* **58**, 2387–2389 (1987).
12. Procaccia, I. in *Chaos 87* (ed. Duong-Van, M.) 527–538 (North Holland, Amsterdam, 1987).

Itamar Procaccia is in the Department of Chemical Physics, Weizmann Institute of Science, Rehovot 76100, Israel.

Vision

The eye of the mantid shrimp

R.C. Hardie

ONE of the most engaging, if pugnacious, groups of crustaceans are the stomatopods, the so-called mantid shrimps. Renowned for the impressive power of their raptorial appendages, they are also owners of an unorthodox pair of compound eyes. The most obvious peculiarity of these eyes is their concave construction which allows different regions of the same eye to converge on the same point in space. Although this has been known for almost a century, on page 557 of this issue[1] Justin Marshall describes an unsuspected degree of specialization in the design of a narrow band of medial ommatidia which, in principle, could allow for parallel processing of spatial, spectral and polarization cues.

The four modern stomatopod families, which are believed to have split off the main crustacean line about 400 million years ago[2], have evolved into formidable predators. A lethally designed pair of forward appendages is used to strike prey or conspecific competitors in a manner reminiscent of the praying mantis. The strike is awesome in its power, and with impact velocities of 10 metres per second, achieved within 4–8 milliseconds, represents one of the fastest movements in the animal kingdom[3].

No less remarkable are their eyes. Although these are of typical apposition compound-eye construction — that is, consisting of numerous ommatidia, each monitoring one point of the image — they are cylindrical in shape and characterized by a central groove or midband consisting of between two and six rows of enlarged ommatidia[1,4,6]. The eye is optically (and in some species, physically) concave about the central groove, so that up to three different regions of the eye simultaneously monitor the same point in space[4,6] (Fig. 1). This eye thus has the potential for 'binocular' — or even trinocular — vision with a single eye[4], an ability which may be particularly useful for triggering the strike. The shrimps' namesakes, praying mantids (whose strike, incidentally, is positively lethargic in comparison), have a range-finding mechanism based on stereopsis between the two eyes[5]. Like most other decapods, however, the eyes of mantid shrimps are mounted on mobile stalks (Fig. 1 inset). Because of the lively and largely independent movements of these eyes, a range-finding mechanism based on two eyes may be prone to error[6]. The unusual development of a concave compound eye can thus be seen as an adaptation to provide monocular stereopsis[4,6].

This hypothesis of stomatopod eye function was suggested in 1891 by Sigmund Exner[4]. Marshall's new anatomical study of the ommatidia in the midband region of the eye reveals an even more surprising organization. In the gonodactylid species he investigates, each of the six rows of the

Fig. 1 The Californian gonodactylid *Hemisquilla ensigera*. The strike of this species, which can reach lengths of 25 cm, is capable of breaking aquaria walls. Three black streaks, the middle of which lies in the midband region, can be seen in one of the prominent cylindrical eyes (inset). These are the so-called 'pseudopupils' and represent the three regions of the eye looking at the camera. (Photo courtesy of Professor Malcolm Burrows.)

midband — all of which look in the same direction — has its own characteristic structure. Two rows (1 and 4) are of fairly orthodox construction, but the others show unexpected specializations.

In rows 5 and 6, the rhabdomeres (Fig. 2) have very specific microvillar orientations, for example, the rhabdomere of the most distal photoreceptor (R8) in row 5 has a microvillar orientation orthogonal to that of R8 in row 6. Because arthropod photoreceptors are sensitive to polarized light, with maximum sensitivity when the light is polarized parallel to the microvilli, such arrangements may be expected to be involved in the analysis of polarized light.

As far as is known, the specializations of rows 2 and 3 are unique among arthropods, although strikingly analogous to oil droplets in vertebrate cones[7] (Fig. 2). At two discrete levels (once near the top and once half-way down the rhabdom), the photoreceptors no longer produce microvilli, but form a dense collection of vesicles containing carotenoid pigments. In cryosections, this arrangement results in the appearance of up to four differently coloured filters in combinations characteristic for each species (Fig. 3). Again, a possible behavioural role can be identified. Many gonodactylid species are characterized by brightly coloured spots (meral spots) on their forward appendages. These are conspicuously used in display and their colour is often the only obvious distinguishing feature between different species[2]. In another family of mantid shrimps (squillids), the meral spots tend to be dull in colour[2] and there is no indication that their rhabdoms have such filter blocks[8].

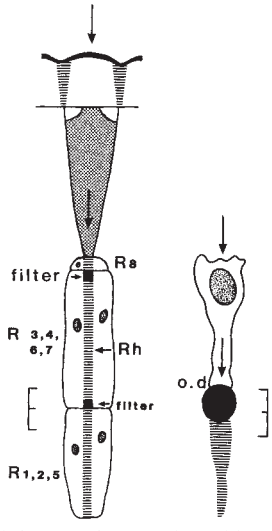


Fig. 2 Oil droplets in gonodactylids and vertebrate cones. Left, a single ommatidium from row 2 or 3 of the midband region. Two discrete filter blocks are incorporated into the light-guiding rhabdom (Rh) which is made up of densely packed microvilli from the eight photoreceptor cells (R1–8). Each cell's contribution to the tiered rhabdom is called a rhabdomere. Bar, 100 μm (adapted from ref. 1.) Right, vertebrate cone with oil droplet (o.d.) Bar, 10 μm.

The filters will inevitably cause marked shifts in the spectral sensitivities of the photoreceptors, and the intriguing possibility, suggested by Marshall, is that the different filters could be used to construct a colour-vision system based on a single photopigment. This idea has often been suggested, particularly in relation to oil droplets in vertebrate cones[9]. But, with the exception of a rather crude dichromatic version believed to operate in the eye of a cricket[10], a concrete example of this attractive hypothesis has yet to be found. The omens are quite promising, as most crustaceans studied to date, including the related *Squilla* (Squillidae), appear to have only a single photopigment in their main rhabdom[8]. Microspectrophotometry of the visual pigments and behavioural analysis are now required to determine whether a single photopigment colour-vision system is in fact realized in the gonodactylids.

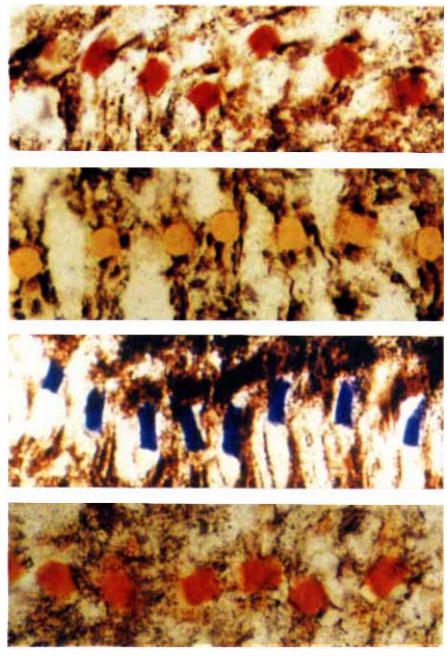


Fig. 3 Cryosection of electron-dense oil droplet structures in mantid shrimps. These filter blocks appear red, yellow, orange, pink, blue and purple, and are likely to contain carotenoids. They are situated at two discrete levels in all rhabdoms of R2 and R3 of the midbands — F1 and F2 (see Fig. 2 of ref. 1). The combination of colours is species-specific and may relate to body coloration. (See ref. 1; photo courtesy of Justin Marshall.)

Whether or not this turns out to be the case, the most significant feature of the gonodactylid eye is the remarkable degree of parallel processing of the retinal image. Mantid shrimps seem to use different ommatidia in the six rows of the midband to gain information about the colour and polarization of single points on the object, while other ommatidia above and below the midband may be used to estimate the object's distance by triangulation. □

1. Marshall, N.J. *Nature* **333**, 557–560 (1988).
2. Caldwell, R.L. & Dingle, H. *Scient. Am.* **234**, 80–89 (1976).
3. Burrows, M. *Z. vergl. Physiol.* **62**, 361–381 (1969).
4. Exner, S. *Die Physiologie der facettirten Augen von Krebsen und Insekten* (Deuticke, Leipzig, Vienna, 1891). (New edn in English, Springer, Berlin, in the press.)
5. Rossel, S. *Nature* **302**, 821–822 (1983).
6. Horridge, G.A. *Endeavour* **1**, 7–17 (1977).
7. Bowmaker, J.K. *Trends Neurosci.* **3**, 196–199 (1980).
8. Cronin, T.W. *J. comp. Physiol.* A**142**, 199–202 (1985).
9. Hailman, J.F. *Nature* **204**, 710 (1964).
10. Kong, K.-L., Fung, Y.M. & Wasserman, G.S. *Science* **207**, 783–786 (1980).

R.C. Hardie is in the Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK.

Immunology

Intolerable secretion in tolerant transgenic mice

Peter Parham

"It is anticipated that the expression in transgenic mice of cloned gene segments, under control of tissue-specific promoters, will facilitate dissection and eventual elucidation of the elaborate interplay between environment and genome that constitutes autoimmune disease". This opening sentence from my gestating grant application propounds the hopes, the party line. In reality, transgenic mice are the mystery tour of immunology; they are new and exciting and no one knows where they will go. This is vividly illustrated by the paper of Allison *et al.* on page 529 of this issue[1]. These authors set out to establish a simple mouse model for diabetes, caused by immune destruction of pancreatic β-cells. Instead they have come up with a new form of the disease, resulting from some strange cell-biological asphyxia, and novel insights into immunological tolerance. Perhaps this seems bizarre, but in fact similar results were obtained in two related investigations[2,3].

Autoimmunity and diabetes

Susceptibility to many diseases is linked to class II genes of the HLA region; the major histocompatibility complex (MHC) of humans, an example being insulin-dependent diabetes mellitus (IDDM) in which *HLA-DR3* and *DR4* alleles correlate with a compounding susceptibility and the *DR2* allele with resistance. Many HLA-associated diseases are thought to be caused by autoimmune attack on specialized tissues. In the case of diabetes the targets are the insulin-secreting

β-cells of the pancreatic islets. Study of animal models implicates helper T cells in the disease, and in diabetics the destruction of β-cells is associated with autoantibodies and a pancreatic infiltration of B and T lymphocytes[4].

The dual involvement of class II MHC and helper T cells fits well with our current understanding of the immune response. For it is the binding of processed fragments of proteins by class II MHC molecules and their presentation to helper T cells that initiates antigen-specific immunity. Logic then argues that IDDM involves helper T cells responding to peptides derived from insulin and other β-cell-specific components that are presented by class II MHC.

In the normal course of events a state of immunological tolerance prevails, preventing destructive responses to self. What unusual sequences of events can cause the breaking of tolerance and the start of an autoimmune disease? To answer this question requires knowledge of the autoimmune response, the target antigens and the mechanism by which tolerance to those components is normally maintained.

T-cell receptors interact with a composite ligand involving a peptide bound to either a class I or a class II MHC molecule. The repertoire of specificities is moulded in the thymus by selection of receptors stochastically arising from rearrangements of the α, β, γ and δ-chain genes. Current models propose this education of thymocytes to involve a primary course of positive selection followed by a secondary course of negative selection (Fig. 1). Presumably, the templates for selection involve complexes of MHC molecules with peptides derived from proteins that are found in the thymus. Tolerance to self MHC and self peptides results from the negative selection and is proposed to occur by either paralysis, suppression or elimination of T cell clones. For one MHC molecule (I–E) and one non-MHC component (Mls^a) the latter mechanism has been shown to be operative, leading to detectable holes in the T-cell repertoire[5].

Models of thymic selection do not adequately explain tolerance to tissue-specific proteins that are not in the thymus. Such proteins, which include insulin, thyroglobulin and immunoglobulin, can become the targets for autoimmune responses, providing evidence for circulating T cells with specificity for these proteins and against clonal deletion as their mechanism of tolerance.

As an alternative to manipulation of the T-cell repertoire, tolerance to tissue-specific products can arise by preventing MHC expression and antigen presentation. Disease could then arise through aberrant MHC expression — an idea supported by observed changes in MHC expression. Although pancreatic β-cells of healthy individuals have no class II MHC molecules and low levels of class I, the β-cells of the diabetic pancreas express high levels of class I and II MHC. However, the relationship of cause and effect between expression of class II MHC antigens and diabetes as assessed by lymphocyte infiltration has engendered controversy[6]. One view is that aberrant class II expression results in presentation of β-cell antigens, stimulation of specific T-helper cells and lymphocyte infiltration. Cytokine-induced increases in class I MHC expression could then lead to cytotoxic T-cell destruction of the β-cells. The alternative scheme is that lymphocyte infiltration and an ongoing immune response to the β-cells (possibly initiated by viral infection) precedes elevation of MHC expression, which is a secondary effect of the cytokines and interferon-γ produced by the infiltrating T cells.

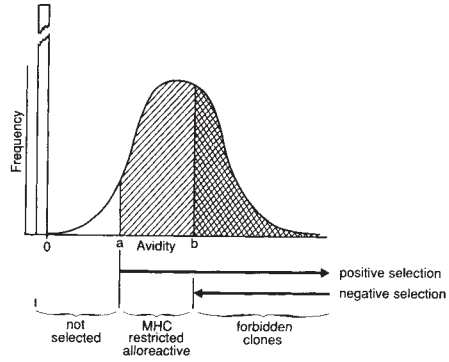


Fig. 1 Hypothetical scheme for avidity selection of T cell clones in the thymus. The frequency distribution of receptor avidity for a given MHC molecule is shown. If the receptor repertoire is intrinsically biased towards interaction with MHC, then the number of clones with no avidity will be lower than if there is no such bias. *a* and *b*, Threshold avidity for positive and negative selection, respectively. Clones surviving both selections are found in the area between these two values.

Tolerance in transgenic mice

The goal of the experiments reported by Allison *et al.* in this issue[1], by Sarvetnick *et al.*[2] and by Lo *et al.*[3] was to separate these variables and to determine if enhanced MHC expression in murine pancreatic β-cells is sufficient to cause lymphocyte infiltration, autoimmune destruction and diabetes. Each group chose to manipulate expression of a different MHC locus: Allison *et al.* the class I molecule H–2K^b; Lo *et al.* the class II molecule I–E^b; and Sarvetnick *et al.* the class II molecule I–A^d. These genes were placed under control of the insulin promoter and used to make transgenic mice. As expected, there was strong expression of the transgenes in the targeted β-cells and little expression in other tissues although Lo *et al.* find some expression in the kidney tubule[3]. In contrast, the biological effects of this expression were quite unexpected. As the investigators undoubtedly hoped, the mice became diabetic and suffered a variable degeneration of their β-cells. However, and to their astonishment, they could find no evidence for autoimmunity.

Histological analysis by Allison *et al.*[1] reveals no lymphocyte infiltration associated with β-cell degeneration. Nor does neonatal thymectomy perturb the course of disease, arguing against a pre-eminent role for T lymphocytes. Perhaps their most surprising result is that compatibility between the transgene and the H–2 type of the mouse, including the thymus, is of no consequence.

A similar result was reported by Lo *et al.*[3], who directed expression of I–E^b molecules in the β-cells of mice that do not express I–E molecules. Again, there was no histological evidence for immune reactions in the pancreas. When these authors examined this state of unresponsiveness, neither peripheral T lymphocytes from lymph nodes nor maturing T cells from the thymus showed any respose *in vitro* to I–E^b. This tolerance was specific as both cell populations responded well to class II molecules of another haplotype.

It was then possible to determine if this tolerance to I–E^b results from clonal deletion. In strains of mice with conventional expression of I–E molecules in various tissues, including the thymus, there is deletion of T cells that express receptors containing chains from the $V_\beta17$ family[5]. This phenomenon provided the first example of tolerance through clonal deletion and presumably occurs because these receptors have too high an affinity for I–E molecules (Fig. 1). In contrast to normal mice, Flavell and colleagues find that the I–E^b transgenic mice retained a normal complement of $V_\beta17$-expressing T cells[7]. Tolerance does not derive from deletion and is presumably a result of paralysis or suppression. This result provides an indication that different mechanisms may be involved in maintaining tolerance to intra- and extra-thymic components. The difference in mechanism also argues against the tolerance being due to significant but undetectable levels of the transgene in the thymus.

Tolerance to I–E^b of T cells in both the peripheral tissues and the thymus of the transgenic mice indicates that there is communication of the antigens in the periphery back to the thymus. One mechanism is that the spectrum of molecules from all tissues is transported to the thymus. This seems improbable. A second requires that T cells, which are stimulated in the periphery during development, 'home' back to the thymus to initiate clonal paralysis or the selection of clone-specific regulatory cells.

Generalization from the results of Allison *et al.*[1] and Lo *et al.*[2] may be premature as others have found that similarly constructed I–A transgenic mice, with

lower levels of expression on β-cells, are neither diabetic nor tolerant to the transgenic product (C. Benoist, personal communication). In addition, Arnold *et al.*[8] made mice transgenic for expression of a soluble class I alloantigen (H–2K^k) and found no evidence for tolerance to cell surface-associated H–2K^k.

Diabetes in transgenic mice

If, for once, the immune system is not to blame, then what causes diabetes in these transgenic mice and is it the same in all three systems[1-3]? Common to the three groups of mice studied is the reduction in secretion of active insulin. However, they differ in the histological appearance of the pancreas. For the H–2K^b transgenic mice fewer islets depleted of β-cells, and with disorganized structure, are reported. In contrast, islet tissues in the I–E transgenic mice remain histologically intact, even in animals with severe diabetes[3]. The I–A animals were intermediate[2]. Although no controlled histological comparison has been done, these observations suggest that cell death can be separated from and is not an inevitable consequence of the effect upon insulin production.

To assess the potential sites for disruption, I shall consider the steps leading to formation of insulin. An effect at the level of gene regulation is possible and would derive from competition between the MHC and insulin-associated promoters for transcriptional factors. The experiments of Sarvetnick *et al.*[2] exclude this as a significant contributor. Transgenic mice expressing either the α- or β-chain of I–A were healthy or exhibited only a mild hyperglycaemia, whereas their progeny expressing both α- and β-chains were decidedly sick. This result argues strongly for an effect caused by an excess of MHC protein and perhaps of functional molecules made up of α- and β-chains.

The most likely cause of the transgenic diabetes is through disruption of the pathway for biosynthesis, maturation and secretion of insulin. This could involve non-specific competition for protein and other components that contribute to this process or a specific interaction between MHC and insulin involving Bjorkman's groove: the peptide-binding site of the MHC molecule[9].

Insulin is synthesized on membrane-bound ribosomes as preproinsulin. With translocation to the endoplasmic reticulum and removal of the signal peptide, proinsulin is formed, which subsequently follows a pathway of regulated secretion. On passage through the Golgi stack, proinsulin begins to condense into material that densely stains in the electron microscope. It is not certain if this just involves insulin aggregation or if proteoglycans or other agents contribute. On reaching the trans-side of the Golgi, further condensation occurs at clathrin-coated regions of membrane. Insulin-containing coated vesicles with lowered *p*H are formed and are the compartment where the connecting peptide (C) is proteolytically removed to yield biologically active insulin. This event appears to initiate clathrin depolymerization allowing maturation of the vesicle into a secretory granule[10] (Fig. 2).

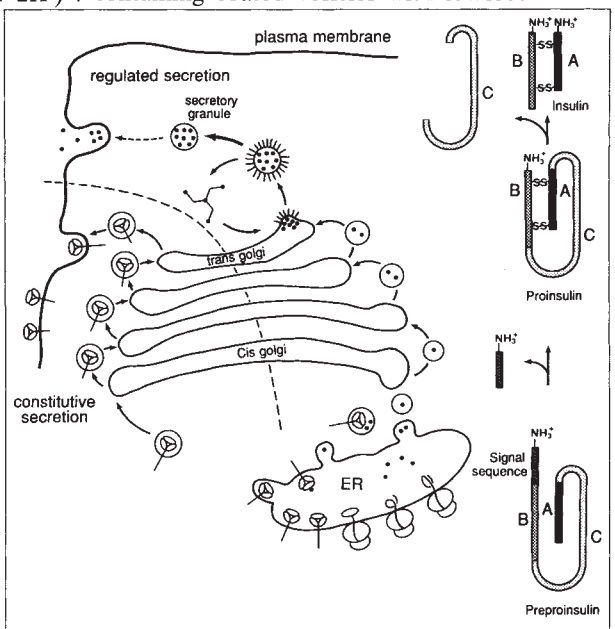


Fig. 2 Pathways of regulated and constitutive secretion. Biosynthesis and exocytosis of insulin (●) by the regulated pathway is shown to the right and that of a class I or class II MHC molecule ( ) is shown to the left. Right, a cartoon of the insulin molecule and its maturation by proteolytic cleavage. A and B, the two chains of mature insulin; C, connecting peptide.

The purpose of the pancreatic β-cell is to make insulin and its biosynthetic machinery is likely to be adapted to this purpose. MHC molecules are membrane glycoproteins that share initial parts of the pathway with insulin and then diverge in the Golgi. Competition for factors that promote the correct folding of newly synthesized proteins in the endoplasmic reticulum could result in the production of inactive insulin. These factors might include protein disulphide isomerase and molecules of the hsp 70 heat-shock protein family[11]. Malformed insulin may be targeted for intracellular degradation or proceed through the secretory pathways with diabetes resulting either from secretion of inactive insulin and/or from disruption of some step in the orchestration of secretion.

A role for Bjorkman's groove?

The cell-biological instincts are to dispense with the matter at this point: the immunologist pursues specificity, a role for the MHC binding site. A direct test of this hypothesis would be to target increased expression of other membrane glycoproteins to the pancreatic β-cells. Would CD4, CD8, Thy-1 or a viral glycoprotein induce diabetes? Although MHC molecules are generally considered to bind peptides derived by proteolytic fragmentation of larger proteins, insulin provides an obvious target. There are cases, like fibrinogen, in which intact proteins are bound through an epitope in an unstructured region[12]. The endoplasmic reticulum contains recent products of the ribosome that are partially folded and perhaps inviting to MHC binding: a reaction that may be enhanced if overproduction of MHC delays its exit from the endoplasmic reticulum. That partially folded insulin may bind to MHC is consistent with experiments showing that a fragment containing about 60 per cent of the mature molecule is presented to T cells by I–A molecules[13].

Individuals have a small number of distinct MHC molecules and a large number of T cells capable of making a response to almost any protein antigen[5]. As MHC presents peptides to T cells it is inevitable that every MHC molecule can bind at least one peptide from almost any protein. Thus for I–A^d, I–E^b and H–2K^b all to bind some bit of insulin is plausible, as is the possibility that these bits will be different, for this is the effect of MHC polymorphism. The difference in the epitope specificity or the strength of binding might define the histological prognosis.

MHC–insulin binding

MHC binding to insulin could prevent delivery of active insulin to secretory granules in various ways: (1) correct folding of proinsulin in the endoplasmic reticulum could be disrupted, leading to an inactive product; (2) condensation could provide the positive signal for sorting along the regulated pathway to the secretory granule. Impairment of condensation in the transgolgi might by default route proinsulin along the constitutive pathway to the plasma membrane or external milieu where it will be retained in the inactive form; and (3) binding of MHC to proinsulin may sterically inhibit the protease that converts proinsulin to insulin. This would result in storage and secretion of an inactive product.

There are no data so far that distinguish between these possibilities. Allison *et al.* show[1] reduced quantities of insulin in the pancreatic islets of transgenic mice and also reduced secretion in response to glucose. At first glance, this might eliminate the last possibility. But in the presence of hyperglycaemia, one expects β-cells to be chronically stimulated to secrete and thus be depleted of insulin.

The binding of MHC might be not to

insulin but to some other protein that is involved in its biosynthesis and exocytosis, candidates being protein disulphide isomerase, hsp 70, the sorting receptor or the protease that cleaves proinsulin. There is a precedent for one protein's changing the pathway of exocytosis and maturation of another. In this case it is the class I MHC molecule that falls victim to the E3/19 protein during adenovirus infection. The complex stays in the endoplasmic reticulum, preventing cell surface expression of class I MHC and presentation of virus-derived peptides to the immune system of the host. The binding of E3/19 to HLA-A,B,C molecules shows polymorphism which maps to the α_1 and α_2 domains[14]. Reminiscent of MHC restriction, this is surely not chance and encourages the hypothesis that a relatively flexible, unstructured part of E3/19 is binding into Bjorkman's groove.

The binding site of MHC molecules lacks the singular specificity of an antibody molecule. It is more analogous to that of a proteolytic enzyme such as trypsin, which is capable of binding and clipping most proteins but shows sequence preferences in the choice of sites. Given the broadness of specificity one should expect and be looking out for additional molecules that perturb, regulate or exploit MHC molecules by latching on to their binding sites. This mechanism could provide the basis for all the non-immunological functions attributed to MHC molecules.

An example, which is doubly relevant to this discussion, is the association of class I MHC antigens with the insulin receptor[15]. Experiments from many laboratories show reciprocal precipitation with antibodies against these two molecules and differences due to polymorphism in the class I MHC molecules. Could this effect, which is not found with class II MHC molecules, contribute to the dramatic loss of β-cell viability in the H–2K^b transgenic mice? Uncontrollable up- or down-regulation of functional receptors for insulin or related growth factors by increased class I MHC expression may perturb the metabolism of β-cells to the point where they become non-viable.

β-cell death

The degeneration of β-cells in these transgenic mice is a variable and relatively slow process, suggestive not of an acute toxicity but of metabolic handicap that gradually wears the cells down. In searching for guidance one immediately thinks of the clathrin-minus mutants of yeast. Although not always lethal, this mutation disrupts the organization of membranous organelles and severely increases the generation time. Coordinated metabolism is incapacitated and away from the rich media and intensive care of the laboratory these cells would surely give up their life without clathrin[16]. And it is in the recycling and marshalling of clathrin pools in the pancreatic β-cells that there might lie some clues. Orci and co-workers show that cleavage of proinsulin to insulin occurs in the coated vesicles that bud off from the transgolgi[10]. Prevention of proteolysis by either monesin treatment or introduction of non-cleavable analogues of lysine and arginine into insulin also stop disassembly of the clathrin coat.

These authors hypothesize that generation of the C peptide provides the signal for clathrin disassembly. I have already discussed how binding of proinsulin by MHC could prevent conversion to insulin. Another possibility is for MHC to be incorporated into the coated vesicles and bind the connecting peptide once it has been cleaved. In both situations clathrin would accummulate; locked up in its own cage and unavailable for other functions. Even a partial reduction in receptor-mediated endocytosis could spell disaster for cells and the eventual cause of death could be as unrelated to histocompatibility as the deprivation of iron.

From these[1-3] and from other[5] studies it is apparent that transgenic mice will be instrumental in unravelling thymic education and the questions of tolerance. In contrast to previous experimental systems, they allow introduced components to be present throughout ontogeny. It was perhaps the unanticipated importance of this difference that led Allison *et al.*[1] and Lo *et al.*[2] to be unprepared for the results they obtained. The value of transgenic mice for research on, if one can say it, 'normal' diabetes is at best unclear. Unfortunately, these transgenic mice have not delivered a valid test for the role in this disease of increased expression of MHC molecules by pancreatic cells. However, study of these mice should further understanding of exocytosis and secretion and perhaps the perturbations caused by intracellular MHC molecules. □

1. Allison, J. *et al. Nature* **333**, 529–533 (1988).
2. Sarvetnick, H., Liggitt, D., Pitts, S.L., Hansen, S.E. & Stewart, T.A. *Cell* **52**, 773–782 (1988).
3. Lo, D. *et al. Cell* **53**, 159–168 (1988).
4. Eisenbarth, G.S. *New Engl. J. Med.* **314**, 1360–1366 (1986).
5. Robertson, M. *Nature* **332**, 18–19 (1988).
6. Baird, J.J., Bone, A.J., Walker, R. & Cooke, A. *Nature* **329**, 493–494 (1987).
7. Flavell, R.A. *et al. 6th HLA/H2 Cloning Workshop. Abstr. 51* May 18–22 (Airlie, Virginia, 1988).
8. Arnold, B. *et al. Proc. natn. Acad. Sci. U.S.A.* **85**, 2269–2273 (1988).
9. Bjorkman, P. *et al. Nature* **329**, 506–512 (1987).
10. Orci, L. *et al. Cell* **51**, 1039–1051 (1987).
11. Pelham, H. *Nature* **332**, 776–777 (1988).
12. Allen, P.M. *Immun. Today* **8**, 270–273 (1987).
13. Naquet, P., Ellis, J., Singh, B., Hodges, R.S. & Delovitch, T.L. *J. Immun.* **139**, 3955–3963 (1987).
14. Burgert, H.-G. & Kvist, S. *EMBO J.* **6**, 2019 (1987).
15. Fehlman, M. *et al. Proc. natn. Acad. Sci. U.S.A.* **82**, 8634–8637 (1985).
16. Rothman, J.E. *Nature* **319**, 96 (1986).

Peter Parham is in the Department of Cell Biology, Stanford University School of Medicine, Stanford, California 94305, USA.

Daedalus

Blow up the Earth!

TUNNELLING is a slow, hazardous and very costly business, as demonstrated by the proposed Channel Tunnel between Britain and France. And yet, muses Daedalus, it could be so easy. The Earth's crust is actually quite plastic. The roof of a coal-mine tunnel, for example, slowly subsides; its floor, freed of its previous overburden, comes up. An abandoned tunnel is soon squashed flat. So Daedalus plans to reverse this geoplastic process. His idea is to adapt standard oil-well techniques to horizontal drilling, so as to bore a full-length pilot tunnel a few centimetres across. He will then apply an internal pressure, and simply pump the tunnel up to the required diameter.

This appealing scheme faces several difficulties. Most tunnels and boreholes leak; groundwater gets into them. In Daedalus's pressurized tunnel, the hydraulic fluid will leak out. His pilot bore will have to be filled initially with some sort of muddy grouting, pumped in till the leaks are sealed. And although many surface rocks, like clay and chalk, are fairly plastic at room temperatures, harder ones like granite only deform at a useful rate when heated. Because their creep rate typically goes up tenfold for every ten-degree Celsius rise in temperature, quite modest heating should suffice.

So DREADCO's engineers are working on a plan to thread an electric cable through the pilot borehole, fill the hole with water, seal it and pass a heavy current. The trapped, heated water will reach and exceed its critical state, attaining many hundreds of degrees Celsius and many hundreds of atmospheres pressure. This neat method of hot pressurization has other advantages. Water greatly weakens all rocks, perhaps by lubricating small cracks. And supercriticial water is a powerful solvent for many rock components, especially silica. Under this supercritical assault, the rock will not only yield plastically; it will dissolve away. Daedalus estimates that a few megawatts of electrical power per kilometre should pump up quite a respectable tunnel in a year or so — neatly, painlessly and relatively cheaply.

The final product, however, will not be the straight, parallel hollow cylinder that engineers dream of. The softer and less loaded areas of the tunnel (near the entrances, for example) will have been inflated into huge caverns long before the tough and overburdened rocks in the middle have expanded to design size. And geological anisotropies will cause much wandering from the central line. Furthermore, all the heat put into the rocks will still be there, and for several years travellers will find the tunnel passage rather a tropical experience.

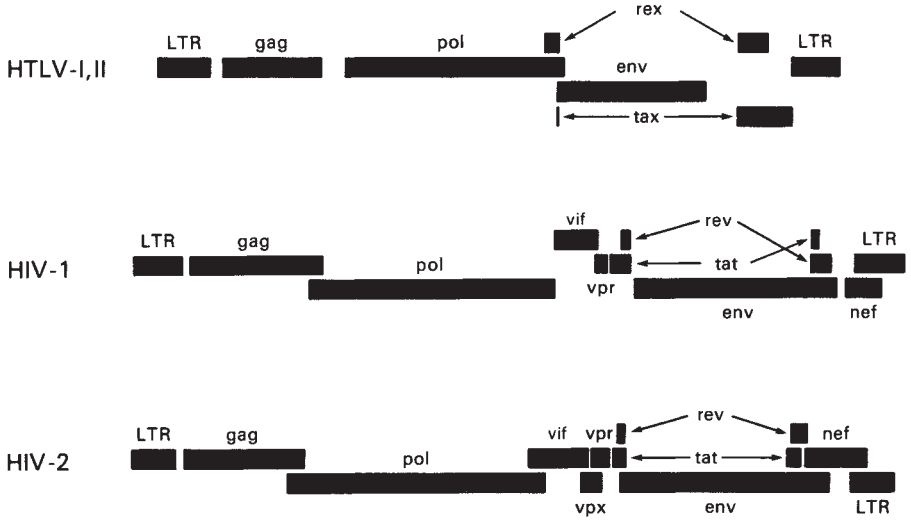
David Jones

HIV/HTLV gene nomenclature

SIR—The complexities of the genomes of human retroviruses (the human T-cell leukaemia viruses, HTLV-I and HTLV-II, and the AIDS-causing human immunodeficiency viruses, HIV-1 and HIV-2) are being unravelled at a rapid pace which is likely to continue and expand. In addition to containing a large ensemble of positive and negative regulatory genes that orchestrate virus expression, these viruses are also remarkable in that they seem to have converged onto parallel regulatory pathways. Two of the regulatory genes of the immunodeficiency viruses are analogous to the two regulatory genes of the leukaemia viruses, although their detailed mechanisms of action may be quite different. Deciphering the modes of action of the regulatory genes of these viruses is crucial to the understanding of their pathogenesis as well as to development of therapeutic agents. Because of the tremendous activity in this field, more than one name has sometimes been given to a single gene and the same name may also apply to more than one gene. In the interest of the many new investigators entering the field for the first time, we feel it is important that we reach a standard nomenclature for all known genes of HIV and HTLV. We propose the scheme outlined in the table.

ROBERT GALLO
FLOSSIE WONG-STAAL
National Cancer Institute, NIH, Bethesda, Maryland 20892, USA
LUC MONTAGNIER
Department of Virology, Institut Pasteur, 75724 Paris Cedex 15, France
WILLIAM A. HASELTINE
Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA
MITSUAKI YOSHIDA
Department of Viral Oncology, Cancer Institute, Tokyo 170, Japan

Proposed name (and derivation)	Previous names	Molecular mass ($\times 10^{-3}$)	Known function
HTLV-I and HTLV-II genes:			
tax_1 (transactivator)	*x-lor, p40x, tat_1*	41, 41, 42	Transactivator of all viral proteins
tax_2	*tat_2, TA*	38	
rex_1 (regulator of expression	*pp27x, tel*	27	Regulates expression of virion proteins
rex_2 virion proteins)		25	
HIV genes:			
tat (transactivator)	*tat-3, TA*	14	Transactivator of all viral proteins
rev (regulator of expression of virion proteins)	*art, trs*	19, 20	Regulates expression of virion proteins
vif (virion infectivity factor)	*sor, A, P', Q*	23	Determines virus infectivity
vpr (R)	*R*	?	Unknown
nef (negative factor)	*3'orf, B, E', F*	27	Reduces virus expression, GTP-binding
vpx (X) (only in HIV-2 and SIV)	*X*	16, 14	Unknown

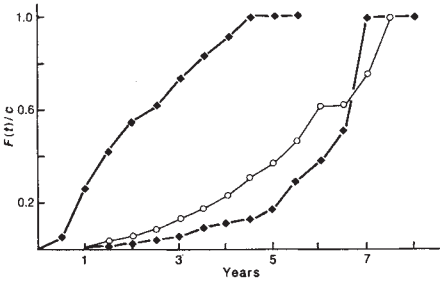


Vpr and *vpx* are temporary names and may be changed when more information about their functions is available. Subscripts 1 and 2 would be used to distinguish genes of HIV-1 and HIV-2 (for example, rev_1 and rev_2). It is expected that genes of the simian viruses (STLV-I, SIV) would follow similar nomenclature with the subscripts STLV or SIV as appropriate.

Estimating the incubation period for AIDS patients

SIR—The nonparametric analyses of the data on transfusion-related AIDS considered by Medley *et al.*[1] indicate problems of identifiability. With data obtained by retrospective determination of the time of infection for diagnosed AIDS cases, it is only possible to estimate the early part of the incubation distribution up to a constant of proportionality. The same applies to the total number of infections by blood transfusion before any given time. The transfusion data themselves are unable to discriminate between high infection rates coupled with long incubation times on the one hand, or low infection rates and short incubation times on the other.

As do Medley *et al.*[1], we postulate a function $h(x)$ which specifies the increase over time of the number of HIV-infected individuals who eventually develop AIDS, and a probability density function $f(s)$ for the incubation time of those individuals. The corresponding likelihood function can be maximized jointly with respect to h and f. As the likelihood depends only on the product of h and f, it is not possible to estimate either of these fuctions completely; they may be individually estimated only up to constants of proportionality c and c^{-1}, respectively. Nonparametric estimates of the proportion of eventual AIDS cases that are diagnosed within t years of infection, $F(t) = \int_0^t f(u)du$, are given in the figure for the three age groups considered by Medley *et al.*. In this figure we show the estimates of $F(t)$ so that for each group, $c = F(7.5)$. For the children, the levelling of the estimate of $F(t)$ by about 3.5 years suggests that the whole of the distribution of incubation times has been seen; it may then be reasonable to suppose that $c = 1$ but, as also noted by Medley *et al.*, a second wave of incubation times that exceed 7.5 years is not excluded by these data. For the other two age groups, there is nothing in the transfusion data themselves to suggest a value for c. As a consequence, it is impossible to place any upper bound on the median incubation time. To estimate this,



Nonparametric estimates of $F(t)$, the probability that the incubation time is less than t for children (◆, top curve), adults (o) and elderly patients (◆, bottom curve). The unspecified constant c is the probability that the incubation time is less than 7.5 years. Note that c cannot be estimated from the transfusion data alone, and that different values of c could apply to the separate age groups.

it is necessary to obtain an external estimate of c.

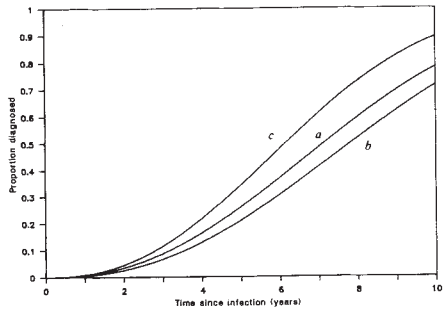
Medley *et al.* comment that estimates of incubation time characteristics such as means and medians are based on extrapolation, and this is true. In fact the difficulties are even more severe. Because of the identifiability problems, even low percentiles of the incubation distribution cannot be estimated without additional assumptions. With parametric assumptions, such as those used by Medley *et al.*, identifiability problems still persist. Thus although point estimates of median incubation times can be obtained, the confidence intervals associated with those estimates are very broad. Under a Weibull model for $f(s)$ and an exponential model for $h(x)$, point estimates (with 95 % confidence intervals) of the median incubation times in years are 2.0 (1.4, 4.0), 7.3 (4.6, ∞), and 5.8 (4.3, ∞) for the young, middle, and elderly age groups, respectively. (Here, ∞ indicates a large but finite upper terminus.) So, not much faith can be placed in the point estimates obtained. Similar remarks apply to the estimation of the total number of infected individuals.

J.D. KALBFLEISCH
J.F. LAWLESS
Department of Statistics and Actuarial Science,
University of Waterloo,
Waterloo, Ontario, Canada N2L 3G1

MEDLEY *ET AL.* REPLY — Kalbfleisch and Lawless are correct in their assertion that if the probability distribution of the incubation period of AIDS is truncated at some time point T (all observations beyond the present time, T, are unobserved), then only the conditional distribution up to T can be estimated without parametric assumption concerning the form of the full distribution of incubation periods. For the matters of concern to epidemiologists and public-health planners, namely understanding and prediction of observed trends, the conditional distribution is of rather little interest and hence parametric formulation is inevitable. We emphasized the obvious hazards of this approach by showing explicitly the different estimates arising from two plausible parametric assumptions (the Weibull and gamma distributions)[1]. We agree that the point estimates of median and mean incubation periods have large errors, and that these are best assessed by likelihood ratio arguments.

In our previous papers[1,2] we noted the importance of updating estimates of the distribution of the incubation period as more data become available. The Centers for Disease Control (CDC) in Atlanta provided a recent data set covering reports of AIDS cases received up to 8 February 1988. In these data there are 560 cases with certain dates of transfusion (infection), diagnosis and report, compared with 297 in the previous data. We have so far considered only those cases who were older than 12 years at diagnosis (512 cases).

In this preliminary analysis we have adopted the same approach as before [1,2]. We assume parametric forms for the distributed incubation period and the distributed reporting lag, and an exponential increase in the incidence of infective transfusions up to early 1985 when routine screening of donated blood was introduced in the United States. We assume that the infection rate has been constant since then. CDC estimate that even after the introduction of screening, about 400 infective donations per year are tested during the period before antibodies against HIV are produced, and therefore appear negative (M. Morgan, personal communication). We are able to discriminate between different parametric forms for the incubation period by comparison of the independent assessment of the rate of infective transfusions with that estimated from the data.



Comparison of the cumulative distributions (Weibull) for the incubation period of AIDS in transfusion-associated patients estimated from: A, all adults (>12 yr) notified to CDC by 8 February 1988; b, people older than 4 yr and less than 60 yr; and c, all cases, regardless of age. Estimates b and c are from a data set in which people were diagnosed before July 1986 and notified to CDC by January 1987. Exact comparisons are not possible because of the difference in age groupings between the two data sets. The most notable features are the similarity of shape between the distributions, and the fact that the means and medians are similar between a and b and between a and c after 2 yr further observation.

We have considered the Weibull and gamma distributions as possible candidates for the incubation-period distribution. The estimated means (medians) are 7.59 yr (7.32 yr) and 24.07 yr (20.84 yr), respectively, and, as previously, the log-likelihood values do not allow a distinction to be made between the distributions. The estimated annual rates of infective transfusion since the introduction of screening, however, are 333 and 2,161, respectively, which leads us to reject the gamma distribution in favour of the Weibull as a descritption of the data.

Also of interest is the observation that the estimate of the overall mean for the adult age classes is similar to that reported in previous publications [1,2]. It is difficult to make direct comparisons as the format of the data has altered. The similarity of the estimates may suggest that the estimate is approaching the true mean of the full distribution, however. Past estimates of this mean have tended to increase with the length of the observation period of transfusion-associated AIDS cases [1-3].

G.F. MEDLEY
R.M. ANDERSON
Department of Pure and Applied Biology,
Imperial College London,
London, SW7 2BB, UK
D.R. COX
Department of Mathematics,
Imperial College London,
London SW7 2BB, UK
L. BILLARD
Department of Statistics,
University of Georgia,
Athens, Georgia 30602, USA

1. Medley, G.F., Anderson, R.M., Cox, D.R. & Billard, L. *Nature* **328**, 718–721 (1987).
2. Medley, G.F., Billard, L. Cox, D.R. & Anderson, R.M. *Proc. R. Soc.* **B233**, 367–377 (1988).
3. Lui, K. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 3051 (1986).

Fish farming and aquaculture

SIR—The article "Fish farming and influenza pandemics" by Christoph Scholtissek and Ernest Naylor (*Nature* **331**, 215; 1988) has been widely reported in Asian and Western newspapers. We feel that it deserves clarification because of adverse effects it might have on the promotion of aquaculture in developing countries.

The authors claim that widespread development of integrated farming systems involving pigs, poultry and fish could lead to the creation of influenza A viruses with new surface antigens against which the human population does not possess specific neutralizing antibodies. The article suggests that pigs may be "mixing vessels" where normally separate avian and human influenza virus reservoirs meet, and where genetic reassortment takes place between them, giving rise to new human pandemic influenza strains.

We share the authors' concern that the

promotion of integrated farming systems in developing countries to alleviate malnutrition should not create potential health hazards. However, we believe that the inferred link between fish farming development in the Third World and human influenza pandemics is grossly overstated.

Pigs and poultry have been brought together without fish on traditional mixed farms in Asia and Europe for centuries, although more recent trends in livestock development in both East and West are towards monocultures because of management and marketing considerations. Most integrated livestock–fish farms raise a single species of terrestrial animal with fish. In the examples cited by the authors, only 16 per cent of the freshwater fish farms in Hong Kong with livestock have more than one species of livestock. Similarly pig–hen–fish culture is not widespread in Thailand and the example discussed in the Commentary (hens in cages above pigs which consume hen faeces) is based on a single farm which has since discontinued the practice in favour of pig–fish integration (the original article implied, but did not actually state, that the practice is widespread — another example of the danger in making a generalization from a very limited data base).

Poultry manure has little nutritional value for intensively fed monogastric animals such as pigs because most of the nitrogen is in a non-proteinaceous form. In short, the co-location of pigs and poultry to supply manure for fish culture is neither prevalent nor likely to become so.

PETER EDWARDS
C. KWEI LIN
DONALD J. MACINTOSH
K. LEONG WEE
DAVID LITTLE
N.L. INNES-TAYLOR
Aquaculture Programme,
Asian Institute of Technology,
PO Box 2754, Bangkok, Thailand

NAYLOR AND SCHOLTISSEK REPLY — In Pullin and Shehadeh (1980), cited in our article, M.N. Delmendo (Food and Agriculture Organization, Bangkok) states "Thailand practices integrated poultry–pig–fish farming, particularly in the central plain where the water supply is abundant". We claimed no more than that pig–poultry–fish systems do occur and pointed out a potential health hazard of expanding the practice of co-locating pigs and poultry around fish ponds, both points which Edwards *et al.* appear to concede.

It would be encouraging if pig–fish systems could be kept separate from poultry–fish systems but to achieve that seems likely to require qualification of a published proposal submitted in February 1988 to the Consultative Group on International Agricultural Research (CGIAR). The proposal being considered by CGIAR argues forcefully for "fully integrated crop/livestock/fish farming with maximal efficiency of use of on-farm resources" and emphasizes the need for a "whole farm" approach to the development of tropical freshwater aquaculture.

ERNEST NAYLOR
School of Animal Biology,
University College of North Wales,
Bangor, Gwynedd LL57 2UW, UK
CHRISTOPH SCHOLTISSEK
Institut für Virologie,
Justus-Liebig-Universität Giessen,
D-6300 Giessen, FRG

Massless particles from a perfect fluid

SIR—Your report[1] on the experiments of Danzmann, Meyerhoff *et al.*[2] on heavy-ion collisions encourages me to draw attention (even if hesitatingly) to a phenomenon which Basilis Xanthopoulos and I[3-5] have encountered in the context of the collision of plane impulsive gravitational waves in general relativity.

Briefly, the phenomenon is this: a perfect fluid, in which the energy-density (ε) equals the pressure (p) and the velocity of sound equals the velocity of light, can be transformed into massless particles, or 'null dust'. More precisely, we showed that the presence of an ($\varepsilon = p$)-fluid in the region of the space-time, after the instant of collision of two impulsive gravitational waves, implies that it is the result of a transformation of null dust following the leading edge of the impulsive waves. This transformation, first encountered in the framework of general relativity, is, as Roger Penrose pointed out to us, equally possible in the framework of special relativity (see ref. 6 and ref. 5 Fig. 1). There can be no doubt that the transformation described does and will take place under the circumstances considered. Indeed, as we wrote, "If the reality of this ultimate state of matter i.e. a fluid with $\varepsilon = p$ is accepted, then the fact that such matter can be converted into null dust, both in the frameworks of general and special relativity, gives credence to the possibility that null dust, as classically defined, may have an equal reality." The principal question concerns the reality of an ($\varepsilon = p$)-fluid.

It was pointed out by Zel'dovich[6] in 1962 that the ultimate equation of state of electrically neutral matter consisting of charged constituents of opposite signs, is one in which equality between the energy-density and the pressure is attained and the velocity of sound equals the velocity of light. (For a simple account of Zel'dovich's arguments see ref. 7.) Zel'dovich estimates that the density at which the ultimate equation of state may be attained is 3×10^{16} g cm^{-3}, that is, ~ 100 times the nuclear density, 3×10^{14} g cm^{-3}. However, already at densities of 3×10^{15} g cm^{-3}, we may expect $p \sim \frac{1}{3}\varepsilon$—the conventional limiting form of the equation of state for a fully relativistic gas.

If we accept the attainability of the equation of state $p = \varepsilon$ as reasonable, how is the null dust described, in the context in which it is transformed into an ($\varepsilon = p$)-fluid? In the framework of general relativity, the Ricci tensor, R^{ij}, is related to the energy-momentum tensor, T^{ij}, by Einstein's equation

$$R^{ij} = -2(T^{ij} - \tfrac{1}{2}g^{lm}T_{lm}g^{ij})$$

with the conventional relativistic units. For an ($\varepsilon = p$)-fluid this relation gives

$$R^{ij} = -4\varepsilon u^i u^j,$$

where u^i denotes the time-like four-velocity of the fluid. Null dust is now described by the relation

$$R^{ij} = \text{constant } k^i k^j$$

where k^i is a null vector. In other words, an ($\varepsilon - p$)-fluid becomes null dust when $u^i \to k^i$, as it does at the shock fronts.

We now turn to the possible relevance of the foregoing remarks to the experiments on the heavy-ion (uranium–thorium) collisions. If we describe the collision as being between two fluids with a density $\rho \sim 3 \times 10^{14}$ g cm^{-3}, then we may expect that at the instant of collision a reflected shock wave results. It is known that the compression following such a reflection can be quite high if the effective 'ratio of specific heats' γ is close to 1. Thus the compression can be as high as 23 even for $\gamma = 1.2$. Because $\gamma \to 1$ even at densities as 'low' as 3×10^{15} g cm^{-3}, it would not appear unreasonable to suppose that a compression exceeding 50 is possible under the conditions of the experiments and that a state of matter approaching Zel'dovich's ultimate state occurs. And if it does, why should we not suppose that conversion to null dust, having the characteristics we have described, takes place?

It does not seem likely that the phenomenon described is relevant to the experiments of Danzmann *et al.*[2] because in their experiments, the energy of the collisions was so arranged that the nuclei just grazed one another. But it may be relevant to experiments of more energetic collisions (apparently now being planned) in which the nuclei will strike one another with such strength that a shock wave of the kind described may indeed result, with the attendant phenomena.

S. CHANDRASEKHAR
The Enrico Fermi Institute,
University of Chicago,
Chicago, Illinois 60637, USA

1. Maddox, J. *Nature* **331**, 109 (1988).
2. Danzmann, K. *et al. Phys. Rev. Lett.* **59**, 1885–1888 (1987).
3. Chandrasekhar, S. & Xanthopoulos, B. C. *Proc. R. Soc.* **A402**, 37–65 (1985).
4. Chandrasekhar, S. & Xanthopoulos, B. C. *Proc. R. Soc.* **A403**, 189–198 (1986).
5. Chandrasekhar, S. & Xanthopoulos, B. C. *Proc. R. Soc.* **A414**, 1–30 (1987).
6. Zel'dovich, Y. B. *Sov. Phys. JETP* **14**, 1143–1147 (1962).
7. Shapiro, S. L. & Teukolsky, S. A. *Black Holes, White Dwarfs and Neutron Stars* (Wiley, New York, 1983).

In search of a miracle

Zhores A. Medvedev

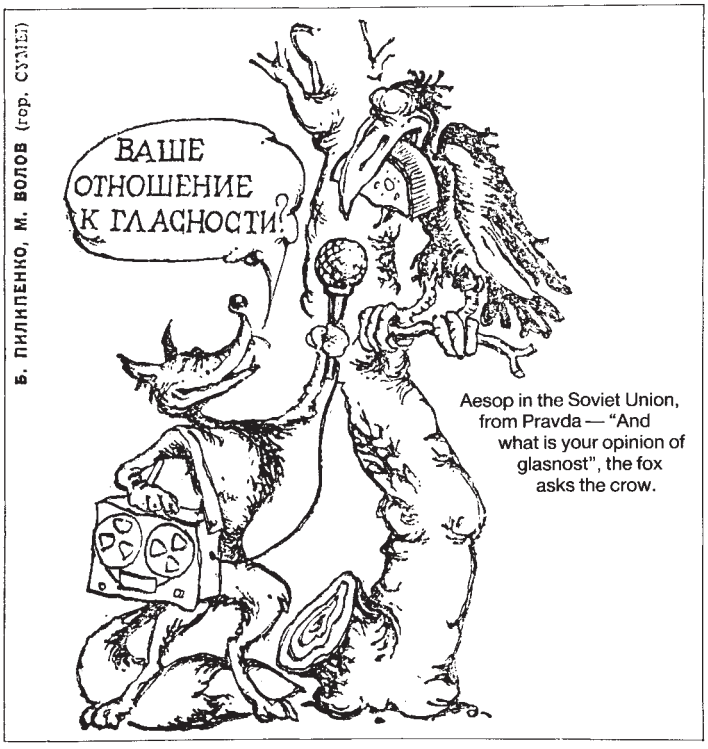
Science and Technology in the USSR. Edited by Michael J. Berry. *Longman: 1988. Pp.405. £63, $95.*
The Status of Soviet Civil Science. Edited by Craig Sinclair. *Martinus Nijhoff:1987. Pp.289. Dfl.195, $96, £60.*

Perestroika in the Soviet Union was originally expected to be concerned not with economic or social reform, but principally with science and technology. The first problem for Gorbachev and his colleagues was the inability of Soviet industry and agriculture to use scientific innovations and new technologies. Nearly half of the world's engineers and research scientists work in the Soviet Union, but the country was clearly not enjoying the fruits of the electronic, computer and biotechnological revolutions.

This picture determined Gorbachev's immediate priorities. The main topics of his first Party Plenum, on 23 April 1985, were linked to accelerating scientific and technological progress. An urgent All-Union Conference was convened two months later and Gorbachev's report there considered developments in science and the modernization of technology to be "the fundamental question of the economic policy of the CPSU". Computerization, application of results of scientific research, innovation and so on were the key parts of this first *perestroika*.

In July 1985, the Central Committee and the government passed a decree which revised the system of salaries for Soviet scientists and technologists. The new regulations did not increase salaries in research establishments, but made them less dependent on academic degrees and more dependent on performance. They also eliminated discrimination against scientists working in industry, whose academic qualifications had not previously been taken into account. There were many other indications that Gorbachev intended to generate rapid scientific–technological change, and his first slogan was *uskorenije* (acceleration), not *glasnost* or democratization as in 1987.

The expectation of a possible Soviet scientific 'miracle' was the subject of a NATO symposium held in September 1986, the proceedings of which are presented in *The Status of Soviet Civil Science*. They also contributed to the decision by Longman to update and revise *Science and Technology in the USSR*, which was originally published in 1976. Both books, however, in fact show not how much Soviet science has changed since 1985, but that it remains essentially the same despite numerous government decrees and party mandates.



Aesop in the Soviet Union, from Pravda — "And what is your opinion of glasnost", the fox asks the crow.

Berry's book was not designed as an analysis of the performance of Soviet science and technology, rather as a guide to the country's enormous scientific establishment. Only three brief chapters describe the history and current position of science in the Soviet political and economic system. The remaining 24 are substitutes for non-existent 'handbooks' of the USSR Academy of Sciences, Academy of Medical Sciences, Academy of Agricultural Sciences, republican academies, numerous ministerial research networks, scientific societies and all other institutions concerned with science and technology.

The reader can determine how big the Soviet scientific system is (5,057 research establishments), how much manpower it employs and how much money it consumes. The most important research institutes, societies and academies are listed with full addresses and names of the senior staff. Most of the professional and popular science journals are also given, with their addresses, affiliations and editors. The task of compiling this information was not an easy one; the authors used, among other sources, a recent edition of the Moscow telephone directory and various city guides. By its nature this is not a readable or thought-provoking book, but it will be of high value to those Western officials in science who are in regular contact with their Soviet counterparts, librarians who have extra funds to spend on Russian-language scientific journals, and science and technology departments of the Soviet Embassies in the West who do not have anything similar in their own language.

While Berry's volume was not intended to be anything but a guide, the NATO symposium was convened in an attempt to analyse the possible consequences of the Soviet scientific–technological revolution. The book is not as comprehensive in its scope as Berry's and does not even include an index. But it is much more interesting to read because the authors try to explain the real problems which face Soviet science in a rapidly changing world and the attitude of Soviet leadership which wants to reform its science and to make it more productive.

Modern science is crucially dependent on computers and electronic communications. The Soviet lag in this area is unavoidable, simply because the Soviet and East European computer industries, which still prefer to copy Western models, have not been able to reproduce the Western communication infrastructure. Personal computers are rare, there is no national data transmission system, the telephone lines are of poor quality and there is no access to foreign databases. Even in Academy institutes, those who have access to computers may not be able to use printers and photocopiers. Censorship of scientific communications is still common — the Soviet research system failed to respond to *perestroika* and *glasnost* in the same way as the general media and literature.

Neither minor changes in the salary scales, without an overall increase in the research and development budget, nor the creation of commissions to encourage cooperation between academic and ministerial research networks, have produced visible results. For decades it did not pay for Soviet scientists or engineers to take risks and to be inventive. The old pay sys-

tem made it easier for ministries to create independent research institutes rather than bureaux directly attached to industry. Inevitably this created a rift between research and production in the different branches of industry. The heavy dependence on salaries from academic qualifications made scientists spend too much of their productive lives in obtaining degrees, not practical results. Eventually science lost its attractiveness for young talent — nearly all prominent Soviet scientists are now well over the age of 50. It was not, therefore, surprising that the participants at the NATO symposium failed to find either *perestroika* or revolution in Soviet science. Only a gradual evolution over the next decade was predicted which *may* make research and development "more responsive to the needs of the country".

This was certainly Gorbachev's own conclusion, after three years of waiting for results, and more radical measures have recently been introduced by the Central Committee. From 1988 all the ministerial research networks (which take about 60 per cent of the Soviet research and development budget, the rest being consumed by the academies and higher-education establishments) have been transferred to 'self-financing' schemes which will enable them to earn money from contracts within industry, not from the state budget. Undertaking business trips to socialist countries has been made simpler; travel to Western countries is also under review, and might be made more flexible. Retirement at 70, even for academicians, is now mandatory. And the problems of young scientists, who have been awarded a 20–30 per cent increase in salary (which had dropped to half that of an average industrial wage during the 1970s), have also been tackled. These and other reforms are all designed to bring about an "acceleration" in the development of science and technology.

Far reaching as such reforms might seem, they remain part of an attempt to prove the benefits of 'socialist' science and to make it a showpiece of the advantages of the Soviet political and economic system. Soviet science has great potential, which is apparent from its excellence in many theoretical fields. But all the current reorganization is essentially bureaucratic and carried out from above. The Soviet leadership still refuses to acknowledge the global characer of scientific and technological development. The best way for the Soviet Union to benefit from Western success in science and technology is more likely to be in integration with it, not in an attempt to generate its own separate revolution. □

Zhores A. Medvedev is at the National Institute for Medical Research, Mill Hill, London NW7 1AA, UK. His most recent book is Soviet Agriculture, *which was published late last year by W.W. Norton.*

Transformation in biochemistry

David Saggerson

Biochemistry, 3rd edn. By Lubert Stryer. *W.H. Freeman: 1988. Pp.1,089. Hbk $47.95, £39.95; pbk £24.95.*

It is seven years since the second edition of this enormously popular book appeared, and the need for overhaul and updating had clearly become pressing. Remaining as the sole author, Lubert Stryer has performed a prodigious feat to produce a worthy successor. To those for whom the second edition had become an old friend, much of the new edition will be familiar. Many of the excellent illustrations are retained, together with a substantial portion of the text. For all that, it is clear that a major transformation has been wrought.

With 1,089 pages of slightly larger size, the increase in information content is about 40 per cent over that in the second edition. The book is now divided into six rather than five parts as a result of the creation of a new Part I entitled "The Molecular Design of Life". This describes the basics of protein and nucleic-acid structure, and outlines the flow of genetic information, gene analysis and DNA cloning methods. The new section therefore reflects the growing interrelationships between biochemistry and molecular biology by showing, in particular, how techniques of protein and nucleic-acid chemistry have become mutually reinforcing. The requirement for an early introduction to some molecular biology becomes obvious when, in later chapters, frequent reference is made to RNA splicing and to the correspondence of exons with functional units within proteins.

The sequence of the following five parts ("Protein Conformation, Dynamics and Function"; "Generation and Storage of Metabolic Energy"; "Biosynthesis of Macromolecular Precursors"; "Genetic Information"; and "Molecular Physiology") is the same as before and remains logical. My only quibble is that photosynthesis arrives only in Chapter 22 and is therefore separated by almost 100 pages from Chapter 17 on respiration and oxidative phosphorylation. I would have thought there was a compelling case for these two topics to be placed side by side.

Striking full colour illustrations continue to be a major feature throughout and colour computer graphics have now also been employed to illustrate proteins and nucleic acids in three dimensions. As before, each chapter is followed by a list of selected reading and a set of problems. The suggested readings run up to 1986 and in many cases go considerably beyond the subject matter covered in the text. I challenged the index on numerous occasions and found it to be thorough and complete.

What individual topics are particularly new and timely, or have been thoroughly reworked in this edition? The following list is not exhaustive, but one should particularly mention blotting techniques; solid-phase methods for peptide and oligonucleotide synthesis; monoclonal antibodies; site-specific mutagenesis; insertion and expression of genes in plant and animal cells; reconstitution of membrane systems; use of nuclear magnetic resonance in metabolic studies; the chemiosmotic hypothesis (a very readable account); protein targeting to specific cellular locations; viruses and oncogenes; and transmembrane signalling processes.

A particular strength of the book is in

High life — Montezuma Castle, Arizona, which is thought to have been occupied by Singua Indians between the thirteenth and fifteenth centuries. The picture is reproduced from David Kempe's Living Underground: A History of Cave and Cliff Dwelling, *to be published by Herbert Press, London, on 23 June. Price is £18.*

Stryer's presentation of the many and varied features of protein structure, and in his detailed but readable account of gene structure and expression. If there is a weakness, it is in the accounts of metabolic reactions and pathways. These are clear and understandable, but I was disappointed in the attempts to integrate various aspect of metabolism, to show how they are controlled and to put all of this into a physiological setting. Although one can glean quite a lot of information about aspects of metabolic regulation from assorted sections of the text, even with a chapter entitled "Integration of Metabolism" one still has largely to form one's own overall synthesis. Apart from providing a good account of hormone mechanisms, the final section dealing with molecular physiology also fails to resolve this problem, because the emphasis is on the 'molecular' rather than the 'physiological'.

Despite my reservations about the treatment of metabolism, overall this is an outstanding book. There is so much competition in the textbook market these days, the publisher could scarcely aim for anything less. □

David Saggerson is Reader in the Department of Biochemistry, University College London, Gower Street, London WC1E 6BT, UK.

The making of the maladapts

Natural Obsessions: The Search for the Oncogene. By Natalie Angier. *Houghton Mifflin: 1988. Pp. 394. $19.95.*

DEAR NATALIE,

Wonderful book! The story blew me away. You say "MGM or Columbia Pictures might have turned it into a film, if Hollywood wasn't so thoroughly convinced that scientists are a group of unphotogenic maladapts". Well, have I got *news* for you. We want to make that movie! I'm really very very juiced up about the idea.

Those scientists you hung out with and wrote about have got real edge. Never mind they work on cancer. Handled right it can pull the crowds in; remember Love Story? Of course we'll need to make some things just a tad simpler. You have to reckon that most movie goers are just a shade sly of completely dumb when it comes to science.

Even I'm more at home with Eisenstein than Einstein, and tho I was concentrating some I didn't understand all you wrote about the oncogenes that freak those guys out at the Whitehead and Cold Spring Harbor. But for the movie, the details won't matter. We'll just need to make it clear that the research will likely cure cancer, and real soon.

You know, you're dead right when you accuse us film makers of stereotyping scientists. I guess we do tend to make them awkward and asexual, with the men having acne and glasses, and the women glasses and loud, whining voices. But we'll change all that.

I can't tell you how crazy casting were about your main characters. Especially Bob Weinberg — "the Doris Day of molecular biology". Keep this quiet, but they will likely offer the part to Richard Dreyfuss. (Could you get hold of a pair of those funny fat-soled shoes you say Bob wears for his flat feet for our costume designer to check out?) I'm not going to say who we have in mind to play Mike Wigler but he's close to your description ". . . a little bit of a schlemiel . . . yet his face has a cuteness to it — when he smiles he looks like a troll doll — and he's too aware of what's going on around him to qualify as a nerd".

There are two incidents that we have slated for the treatment. One is when the Weinberg lab luck out with the metastasis gene and find they've reisolated the ras gene. The other is when Mariano Barbacid comes from nowhere and nearly beats Weinberg and Wigler in getting the ras gene in the first place. What's really great is that he's Spanish — for which, since we've gotten orders to find parts for minority actors, read Hispanic.

And we just *love* those sexy young postdocs. You know, Snezna Rogelj, who "isn't just pretty; she's dazzling . . . in a short leather skirt she made from Indian purses". And Cliff Tabin, "delightfully handsome in a husky, masculine, linebacker sort of way"; René Bernards, who "looks like a Leonardo sketch of a young horseman, or at least a contemporary interpretation by Calvin Klein"; and Cornelia Bargmann, "one of the few almost perfect people I've ever met".

You know what's missing though? Sex. With all those good looking guys working late at night, surely something must have gone on oftentimes. (Otherwise we've been *right* about the asexual stereotype.) Anyway, we're going to need sex; without it the movie will be a bummer.

What about the ending? I guess the story shouldn't finish with the ras genes or the discovery of complementary oncogene types, or even the discovery of the 'neu' oncogene. Hows about the great moment when the Weinberg lab gets the retinoblastoma gene? What do you think? By the way, did you check out the bit where you describe how that Nature editor Newmark wanted to read the retinoblastoma paper before accepting it and then insisted it was published as a short letter? Unbelievable! What a nebbish! Still, we need a bad guy or two.

NEWT P. KRAEMER
Onco Productions,
London and Los Angeles

Hot topics

John L. Roberts

Temperature Biology of Animals. By A.R. Cossins and K. Bowler. *Chapman & Hall: 1987. Pp.339. £32.50, $57.50.*

RANGING from "molecules to the organism, and from physiology to behaviour", the authors of this useful book for advanced undergraduates and postgraduate students provide their readers with a sampling of studies in animal thermobiology. They begin, appropriately, with a discussion of what temperature is — a vexing question for many students — and of brownian movement as a 'thermometer' of kinetic energy. This leads in readable fashion into an account of the physical properties of thermal environments, and to a summary of the search for temperature coefficients that satisfy the physiologist's need for orderly descriptions of temperature effects upon rate processes such as metabolism.

Cossins and Bowler then consider how temperature change affects the lives of bradymetabolic and tachymetabolic animals; here they describe such fascinating variant metabolic patterns as periodic endothermy, a kind of shivering thermogenesis, which is used by dung beetles, honey bees and hawk moths, and lamnid sharks and tuna, to keep locomotory muscles warm. The discussion of biochemical pathways used for aerobic heat production during both shivering and non-shivering thermogenesis makes clear by its brevity that this is a subject deserving of far more attention in the future.

A variety of adaptive mechanisms for heat dumping and heat retention evolved with endothermy, most notably evaporative cooling and the addition of insulating features such as feathers and fur. Just how significant behavioural thermoregulation has been in the evolution of endothermy remains a matter of debate, but evidence supports the idea that neural networks and setpoint controls for thermoregulation already existed when the earliest endotherms appeared. The authors conclude that body size, too, must have been important. This argument finds support in the observation that rates of heat loss from body surfaces, metabolic rates, the power output (and thus heat production) of muscles and the costs of locomotion all scale downward with increasing mass.

A number of ectotherms are considered which show acclimation of rate functions such as locomotion, oxygen uptake, axonal conduction time and epithelial transport, activities that are often modulated by length of the daily photoperiod. The authors then discuss current ideas about how cellular compensations for temperature change occur. These centre

on adaptations of enzyme pathways to substrate usage, selection of specific enzyme isoforms with acclimation, allosteric modulation and homoeoviscous adaptations of membrane lipids related to the degree of lipid saturation that may alter enzyme properties of membrane-bound enzymes. Yet, given the plasticity of response of individual ectotherms to seasonal climatic conditions, the evolution of specific cellular and biochemical modulations was probably conservative. As the authors comment, the overall plasticity of individual ectotherms in adapting to temperature is dependent upon a complex of physiological adjustments; taken together, these adjustments result in dynamic regulatory abilities well beyond those found in experiments with isolated tissues and with individuals examined under controlled conditions.

Another chapter deals with thermal stress effects that induce adaptations in seasonal tolerance limits, freeze tolerance and avoidance, and with the distribution of animals according to multivariate factors such as stress to combined extremes of both salinity and temperature.

The final chapter is concerned with the effects of temperature upon reproduction, development and growth. A just amount of attention is paid to insect development, particularly to seasonal developmental arrest (diapause) which is largely controlled by a kind of integral relating thermal history and the length of the daily photoperiod. The remarks at the end of this chapter emphasize the importance of temperature in the natural selection of regional thermal phenotypes, well adapted to their niches. All of the chapters end with similar summary conclusions that are one of the real values of this short book. □

John L. Roberts is a Professor in the Department of Zoology, University of Massachusetts, Amherst, Massachusetts 01003, USA.

Labelling business

Arthur J. Birch

Organic Chemistry, The Name Game: Modern Coined Terms and Their Origins. By Alex Nickon and Ernest F. Silversmith. *Pergamon: 1987. Hbk £45, $75; pbk £18, $29.50.*

CHEMISTS miss the free flights of fancy, unconstrained by inhibitory realities, which many of their academic colleagues can undertake. With less opportunity, they occasionally play.

John Read once said "A chemist is a poet who has taken the wrong turning". His book *Humour and Humanism in Chemistry* (1947) is a good illustration of this, with gentle, intelligent and sensitive humour, based on scientific relevance, conveyed through rounded stories and poems in English and German. Another classical example of humour was the *Berichte der Durstigen Chemischen Gesellschaft* of 1886, which was so like the authentic one that it is to be found bound with the *Berichte* in the Cambridge University Library. This journal presented surprising scientific insights, for example benzene ring carbons represented by 'blue-headed' monkeys, with their six tails coiled together in the centre — the aromatic 'sextet' about 40 years before Robinson.

This kind of play favourably contrasts with drawing cats' whiskers on aromatic rings, and some of the present book reminds me of the Disney cartoon with a hippopotamus ballet, danced in tutus. It reads like a scrap-book of very variable quality, despite attempts to organize it into themes. The mode of selection of items is hard to discern and there is little to support Read's suggestion of the chemist as a poet. It has a serious purpose in collecting together many given names, but the trivial and the significant are largely indistinguishable. I met at least half of the names for the first time, and do not expect to encounter them again.

I know that I should review this book, and not one yet to be written, but there is a serious gap in the literature concerning chemical names, which have to be invented and should help people in the formulation and expression of ideas. Names devised systematically for computers are so incomprehensible to human beings that they do not assist such creative mental activities, although they permit easy regurgitation of information. As this book shows, the invention of unsystematic names is sometimes rational, based for example on salient aspects of molecular shape, but the results are often fanciful or personal (perhaps, as with a tattoo like "I love Lucy", some of the authors may wish later that they could change their minds).

The emphasis here is that the process is fun, but names should not be invented solely to amuse the inventors or boost their egos. One great mystery in this field, on a par with Fermat's Last Theorem in mathematics, is the origin of 'barbiturate' — did von Baeyer indeed have a close friend named Barbara? This sort of problem is not rationally soluble, but, like the song the Sirens sang, an answer is not beyond all conjecture.

'Trivial' names are usually just that, and it does not matter what they are except short, euphonious and easy to recall. There are, however, other instances of classification where the implications of words are important. Why did the Ingoldian 'electrophilic' and 'nucleophilic' replace the earlier and more correct 'anionoid' and 'cationoid' of Lapworth and Robinson even though all bond-forming reactions involve electrons? Perhaps people had difficulties in remembering which pole is which; perhaps authoritative persistence prevailed. Another example is 'acetogenin' in biosynthesis, which was pushed to replace the earlier and correct 'polyketide'. The latter word indicates immediately the salient feature of the biosynthetic pathway, while the former is misleading and wrong not only on linguistic and scientific grounds but because it runs counter to the accepted convention of 'genin' when used as a suffix. Why therefore was it invented? Perhaps there should be some neutral refereeing system for rival trivial names.

Another facet not adequately examined here, although the case is mentioned, is illustrated by the superb published letter of Jack Edward (transparently disguised as Aloysius S. Smith) about names for directed bonds in hydroaromatic systems, which appeared at a time when a rash of these had broken out following the advent of the useful 'axial' and 'equatorial'. It is necessary to read this letter in full to appreciate its impact. Not before getting half way through does one smell a rat, which then becomes extravagantly odorous. My first suspicion, I recall, was aroused by a reference to Eni Meni and Mini Mo, *J. Hotsitotsi Inst. Japan*. The result was not only to amuse, but to ask the question whether similar 'serious' contributions are equally ridiculous; indeed, it killed by ridicule the quite unnecessary proliferation of names. Very many in this book could go the same way without loss, and some have done so.

Let me, however, not be too grudging. The book is clearly a labour of love, and is well printed and legible. Although I cannot imagine reading it sequentially, one can open it anywhere at random, without bothering much about what comes before or after, and so it is ideal for the bedside. More importantly, the contents underly some significant aspects of the sociology of science (including the rather child-like attitudes of some scientists outside their work), and they are uniquely informative. Here are to be found anecdotal and often unpublished accounts not available anywhere else, especially not in scientific papers. I was intrigued to see a discussion on the origin of 'nor', which I have used for steroids for over 40 years without finding out its origin.

The world of chemical research is a small one and many of its inhabitants will find in this book interesting sidelights on friends and colleagues. Libraries should have it, and organic chemists will, I think, find it interesting to browse in at leisure. I only wish . . . □

Arthur J. Birch is Emeritus Professor in the Department of Organic Chemistry, Australian National University, GPO Box 4, Canberra, ACT 2601, Australia.

The need for national HIV databases

Scott P. Layne, Thomas G. Marr, Stirling A. Colgate, James M. Hyman and E. Ann Stanley

Researchers and public health officials involved in surveying and forecasting the course of the HIV epidemic require complete and unfiltered information from many sources. Governments should respond by establishing national HIV databases.

IN THE United States, the cumulative number of reported AIDS cases is growing as a cubic function of time, and it is estimated that between one and two million people are infected with the human immunodeficiency virus (HIV)[1,2]. The mean period from HIV infection to AIDS diagnosis is estimated to be 8 years[3] and the mean survival time after AIDS diagnosis is about 12 months[4], which indicates that most of those who develop symptoms after being infected with HIV will die within the next 10 years unless medical breakthroughs are made.

The HIV epidemic calls for a rapid and coordinated scientific effort across a broad range of disciplines. To that end, there must be effective mechanisms for the sharing of raw data among researchers and their use to answer several crucial questions — How does HIV transmissibility vary with time after infection and with disease stage? What is responsible for the large variance in incubation time from infection to the onset of AIDS? Will current trends in the spread of HIV infection continue? Can we explain the past rate of growth of the epidemic? Can we predict the most effective ways of slowing the spread of infection? And how extensive must national surveillance programmes be reliably to monitor the prevalence of HIV infection?

Information on HIV is surrounded by sensitive social, legal and ethical issues[5-7], so researchers attempting to answer the above questions do not have access to many important elements of raw data. Finding sources of such data and entering into collaborative relationships happens haphazardly, and formal agreements for sharing data and ensuring their confidentiality are difficult to achieve without a precedent or formal framework. The work of those such as epidemiologists, sociologists and mathematical modellers, who all use raw data from a wide range of fields, is severely restricted by these obstacles.

To reduce the obstacles, we propose the creation of national HIV databases to coordinate the sharing of raw data. At first such databases would address the information needs of researchers in their respective countries, but in time they could be linked internationally. Here we argue for a national HIV database for the United States, but the same principles apply to other countries.

A large number of commercial, educational and government organizations have developed information services on HIV. These services are used by many HIV researchers, but there are still significant gaps for those who require detailed data on social and sexual behaviour and unprocessed epidemiological data on HIV.

Surveillance

The biggest database summarizing epidemiological surveillance for the United States is the *AIDS Public Information Data Set*. This database contains one record for each reported AIDS case and is updated quarterly by the Centers For Disease Control with the following information: age, sex, race, area of residence, month and year of diagnosis, month and year of report, date of the patient's death, primary risk group, additional risk factors, country of birth and opportunistic diseases at diagnosis[8]. To ensure confidentiality, case reports are assigned to six Standard Metropolitan Statistical Areas that have more than one million inhabitants. This large granularity greatly reduces the value of the data. It is not possible to identify the AIDS cases for a particular city nor is it possible to identify diagnosis dates of AIDS cases before 1982. As a result, researchers cannot use these data to analyse the progression of the epidemic for a particular city.

There is, too, a lack of useful raw data in journal articles, government reports and conference proceedings. Most of the data contained in these publications are either highly processed or combined into large blocks of statistical information. As a result, the rapidly growing number of interactive databases and preprinted reports that review the literature on HIV do not help researchers identify useful sources of raw data. Similarly, there are a large number of continuing or completed cohort studies that examine the detailed social and sexual behaviour of groups such as homosexual and bisexual men, haemophiliacs, intravenous drug users, transfusion recipients, heterosexual couples and prostitutes[9]. These studies are the primary source of information on HIV serological status as a function of the rate of practice of specific high risk behaviours and on the amount of sexual contact (or mixing) within a particular risk group. But current information services have not systematically identified these cohort studies nor have they collected details on their individual questionnaires, operational procedures and main findings.

To develop mathematical models of HIV transmission and the progression to AIDS, researchers need several distinct kinds of data sets — they need detailed data from a well-defined cohort study for testing and refining of the model; data from a similar cohort study for validating the predictive value of the tuned model; and, for larger-scale applications of the validated model in epidemiological studies, raw data from a broad range of

Table 1 Example of a standard agreement between users and providers of data

- List the names and affiliations of all the users and providers of data on the agreement
- Obtain prior approval from the providers of data before adding new names to the agreement's user list
- Show no raw data to persons outside the agreement's provider and user list
- Protect the security of the raw data when it is stored outside the database system
- Do not use the names of individuals as identifiers in the raw data subset
- Do not disclose data that would allow identification of individuals by inference
- Submit for review to the providers of data all manuscripts prior to submission for publication
- Cite the providers of data as authors or give them acknowledgement on all manuscripts
- Dispose of all user copies of the raw data held outside the database after completion of research
- Agree to abide by any privacy laws or other special handling restrictions that may apply to the raw data

sources dictated by the particular set of equations in the model.

Regardless of the model's complexity, two general sets of parameters, biological and behavioural, are essential for understanding the HIV epidemic[10,11]. Biological parameters are to do with pathogenesis of the disease and include factors such as variations in transmissibility during the course of the infection, length of carrier state and influence of cofactors (for example, sexually transmitted diseases and age). They are influenced by environment and by factors associated with immune stimulation[12]. Behavioural parameters concern activities that may transmit infection between people and include factors such as level and type of sexual activity, and choice of sexual partners. Because of a scarcity of biological and behavioural information, few of these parameters (of which we give examples only) have been calculated to a precision that is required by the mathematical models of the spread of HIV currently proposed[10]. Nevertheless, it may be possible to obtain better estimates on several of them if more raw data from many different studies were available and were analysed together. For that, it will often be necessary to collate information from a large body of epidemiological and sociological studies.

In its simplest form, a national HIV database would be an annotated directory of raw data. In its most complete form, it would be a storehouse of raw data on HIV infection and the AIDS epidemic. At first, the database would be constructed to allow for iterative design of the data structures; later, it would mature into a high-quality information-management system incorporating the latest techniques in database technology.

Standard agreement

In order to facilitate collaboration between researchers, the national HIV database would furnish a standard agreement governing procedures on sharing raw data (see Table 1). This agreement would be signed by both the users and providers of data, and would apply to a particular set of raw data. The standard agreement would establish a code of conduct among researchers regarding the data's security and privacy, and would save time and energy in establishing collaborations that were based on information supplied by the database. In conjunction with this agreement, we have devised four options for the database that could serve the research community (see the box "Design options for a national HIV database").

Options 1–3 are enumerated according to increasing size of the database and scope of the services that they offer, whereas option 4 offers a decentralized repository managed by a decentralized staff. These four options are not mutually exclusive; rather they offer a range of relationships that the users and providers of data could have within a single database service. Questions to be considered by the HIV research community in choosing a structure for the database service are — Which option best serves the needs of the researchers? Which government agency should oversee development? What is the time scale for construction? Who is responsible for its operation? And what are the costs of construction and operation?

The establishment of a national HIV database will require an extraordinary level of commitment on the part of the research community. Individual researchers and institutions will have to share and protect large quantities of confidential data on the intimate behaviour of individuals. They will also have to share data that could otherwise be hoarded to build their own careers. But such a database is needed — and it is needed soon. □

Design options for a national HIV database

1. An annotated directory of sources for raw data, supplying the descriptive information listed in Table 2 and assisting users in identifying providers. Staff would not assist in establishing formal collaborations between users and providers, so the standard agreement is merely a guideline for sharing data. For more detailed information, each user would have to contact a provider individually; for each request by a user the burden of supplying the data would fall entirely on the provider.

2. Both a directory of sources for raw data and a go-between for the users and providers of data. Staff would arrange formal collaboration between users and providers, and so the standard agreement is an instrument for sharing data. The database service supplies the descriptive information (Table 2) and also assists users in finding data by collecting more detailed information from potential providers. Because the database collects detailed information from providers over time, providers are not repeatedly contacted with preliminary questions from various would-be users of their data, as is the case with option 1.

3. A storehouse of raw data, supplying descriptive information (Table 2), the raw data and the standard agreement for use of that data. According to prior formal arrangements, the database may or may not be obliged to contact a provider before releasing their data to a user. For the provider, the burden of supplying the data is reduced to a single shipment with periodic updates. This database service would require a larger support effort than the above two options, but reduces the need for preliminary discussions between the users and providers of data.

4. A decentralized repository of raw data. Raw data resides at the local, or institutional, level but the resulting information is presented at the national level. Access to the local resource is 'transparent' to the user of the national service. This design gives the local organization a greater degree of autonomous control over the data and how it is structured, but requires standardization of transaction types. Use of database staff and the standard agreement would be handled as in option 3.

Table 2 Example of a database questionnaire for requesting data

- General description of study
- Purpose of study
- Study start and end dates (some studies may be in progress)
- Number of persons investigated
- Criteria for selection of participants
- Basic demographics and descriptions (number of males and females, participant's economic status and so on)
- Location and affiliation of investigators
- Names of investigators involved in data collection
- List of publications generated
- Copy of study protocol
- Copy of all questionnaires used
- Description of raw data formats and content
- Funding source
- Person to contact for further information
- Restrictions that may apply to the use of the data (is it available for a fee, are there copyright protections and so on)
- Information on federal, state or local laws that govern use of data (Privacy Act, for example)
- A copy of the raw data

1. Centers for Disease Control. *MMWR* **36(S-6)**, 1–48 (1987).
2. Morgan, W.M. & Curran, J.W. *Public Health Reports* **101**, 459–465 (1986).
3. Curran, J.W. *et al. Science* **239**, 610–616 (1988).
4. Rothenberg, R. *et al. New Engl. J. Med.* **317**, 1297–1302 (1987).
5. Lewis, H.E. *J. Am. Med. Ass.* **258**, 2410–2414 (1987).
6. Dickens, B.M. *Science* **239**, 580–586 (1988).
7. Walters, L. *Science* **239**, 597–603 (1988).
8. *AIDS Public Information Data Set* (Centers for Disease Control, Jan. 4 1988).
9. *An Inventory of Research Studies Regarding AIDS or HTLV-III/LAV Infection* Section A (Apt Associates, Cambridge, Massachusetts, Jan. 1987).
10. Hyman, J.M. & Stanley, E.A. *Math. Biosci.* (in the press).
11. May, R.M. & Anderson, R.M. *Nature* **326**, 137–142 (1987).
12. Haverkos, H.W. *J. infect. Dis.* **156**, 251–257 (1987).

Scott P. Layne, Thomas G. Marr, Stirling A. Colgate, James M. Hyman and E. Ann Stanley are in the Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.

Chimpanzees and AIDS research

Alfred M. Prince, Jan Moor-Jankowski, Jorg W. Eichberg, Huub Schellekens, Rudolph F. Mauler, Marc Girard and Jane Goodall

Pressure is mounting to relax the regulations on importation of chimpanzees for research. Such a policy is unnecessary and would deepen the plight of an already endangered species.

TWENTY years ago Professor Edward Goldsmith called on scientists not to deplete wild populations of chimpanzees by importing animals for research, but to breed them in captivity instead[1,2]. This viewpoint was subsequently adopted in a joint "Policy Statement on Use of Primates for Biomedical Purposes", issued by the World Health Organization and the European Economic Community in 1982. The statement recommends that chimpanzees and other rare primate species "be considered for use in biomedical research projects only if they are obtained from existing self-sustaining captive breeding colonies". Now there is a new danger to this species, which in the meantime has been so depleted that it is considered threatened in most of its native habitat[3,4].

We have learned of efforts by some in the pharmaceutical industry and by others to persuade governments and WHO that there is a serious shortage of captive chimpanzees for AIDS research. It is alleged that this will impede the development of effective vaccines and therapies to control the disease. The goal is relaxation of the restrictions on importation of chimpanzees from Africa as mandated by the Convention on International Trade in Endangered Species (CITES), to which most Western countries are signatories, as well as a relaxation of the WHO–EEC policy statement on this issue. This was confirmed during a meeting of one of us (J. M–J.) in March of 1988 with the Secretary General of CITES in Lausanne.

Extinction

We believe that further removal of this species from the wild may lead to its extinction. Moreover, if properly managed, the large number of captive chimpanzees in biomedical laboratories and elsewhere is sufficient for the needs of AIDS researchers.

In a study funded by the US Public Health Service (USPHS), the International Species Inventory System (ISIS) of the International Union for Conservation of Nature and Natural Resources determined that, in the United States as of two years ago, there were about 1,200 chimpanzees in biomedical research laboratories and 400 in zoos[5]. Approximately 350 of these are allocated to breeding programmes and are not available for AIDS research. The availability of captive chimpanzees on the market indicates that, in the United States alone, there are about 1,000 chimpanzees which are not listed by ISIS and which are kept in laboratories or by private breeders, or are kept as pets, photographic models or entertainers; many animals kept by private parties eventually find their way into research laboratories when they become too large and unruly to be easily handled. Additionally, some 100 babies are born annually in private and USPHS-funded chimpanzee breeding centres, and this breeding effort could be considerably enlarged if funds were made available. Outside the United States there are 120 chimpanzees at the Netherlands TNO Primate Centre, 150 at The New York Blood Center's Vilab II at The Liberian Institute for Biomedical Research, and about 200 in research facilities in Japan, many of them involved in breeding.

We estimate, conservatively, that half of the 1,670 chimpanzees currently held in research laboratories could be made available for AIDS research if required.

The main need for the chimpanzee in AIDS research is in the evaluation of potentially protective vaccines. To date the most likely candidate vaccines, based on immunization against the envelope proteins of HIV, have failed to protect in the chimpanzee model[6,7]. Basic research is needed to explain these failures, otherwise a long string of abortive chimpanzee trials can be anticipated. Much of the research now being planned will use the HIV-like simian immunodeficiency virus (SIV) in monkey models. Vaccine development studies in the chimpanzee model can involve assessment of the immunogenicity of candidate vaccines to determine whether these give rise to potentially protective neutralizing antibodies, and to assess the effect of dose and adjuvants. Because most candidate vaccines are derived from recombinant DNA or synthetic peptides, there is little reason why this cannot be done in human volunteers as pioneered in the case of baculovirus-derived gp 160 vaccine by the National Institutes of Health. Such studies do not involve exposure to live virus, so they could also be carried out in chimpanzees otherwise reserved for breeding.

Once optimal doses, adjuvants and immunogens have been identified it is then necessary to determine whether immunization is protective. This is most rapidly and economically done by immunizing two chimpanzees and then challenging them with a small dose of live virus to see if infection is prevented. If it is, clinical trials would ensue.

Hepatitis B

Relatively large numbers of chimpanzees were employed in the development of the plasma-derived hepatitis B vaccines for two reasons: first, the immunogenicity work was done before the non-infectivity of the products had been established and therefore it could not be carried out in human beings; secondly, as this vaccine was manufactured from virus-containing plasma, lot by lot safety testing in chimpanzees was required. None of these factors applies to potential HIV vaccines made from recombinant-derived or synthetic peptide sources.

If the above strategy is followed, the number of chimpanzees required for AIDS studies will be relatively small and well within the numbers available in biomedical research laboratories.

Relaxation of the restrictions on trade in wild-caught chimpanzees would have a devastating effect on wild populations, virtually all of which are rapidly approaching extinction. If previous practices are allowed to revive, many more chimpanzees of breeding age will be killed in Africa during attempts to capture their infants for export. From discussions with hunters, exporters and importers we estimate that, on average, ten chimpanzees would die for every infant that eventually reached its overseas destination[8]. Such losses surely cannot be justified. □

1. Goldsmith, E.I. *Nature* **219**, 1292(1968).
2. Goldsmith, E.I. *Science* **162**, 3858 (1969).
3. *Current Threats to the Survival of Chimpanzees in Equatorial Africa* (Jane Goodall Institute, Tucson, Arizona, 1987).
4. *African Primate Red Data Book* (IUCN Conservation Monitoring Centre, Cambridge, UK, in the press).
5. Eckholm, E. *The New York Times* 19 November 1985.
6. Dreesman, G.R. *et al.* in *Retroviruses of Human AIDS and Related Diseases* (eds Girard M., DeThe, G. & Vallette, L. 175–178 (Pasteur Vaccins, Paris 1986).
7. Hu, S.-L. *et al. Nature* **328**, 721–723 (1987).
8. Prince, A.M. in *Trends in Bioassay Methodology: In Vivo, In Vitro and Mathematical Approaches* 81–97 (US Government Printing Office, 1981).

Alfred M. Prince is at the Lindsley F. Kimball Research Institute of the New York Blood Center, 310 East 67 Street, New York, New York 10021, USA, and Vilab II, the Liberian Institute for Biomedical Research, Robertsfield, Liberia. His co-authors are at institutes or laboratories in the United States, the Netherlands, the Federal Republic of Germany and France.

Epidemiological parameters of HIV transmission

Roy M. Anderson* & **Robert M. May†**

* Parasite Epidemiology Research Group, Department of Pure and Applied Biology, Imperial College, London University, London SW7 2BB, UK
† Biology Department, Princeton University, Princeton, New Jersey 08540, USA

Epidemiological data on the main determinants of the transmission potential of HIV-1 in specific at risk groups is slowly accumulating, but many uncertainties remain.

IT is now seven years since AIDS (acquired immune deficiency syndrome) was first recognized[1], and just over four years since the etiological agent, HIV-1 (human immunodeficiency virus type 1), was discovered by scientists at the Pasteur Institute and the National Institutes for Health[2,3]. AIDS is now pandemic throughout the world. There was a 124% increase in the total cases reported between December 1986 and December 1987, and 133 countries reported cases to the World Health Organization in February 1988. Uncertainties surrounded many issues concerning both infection and disease. We need to know the likely demographic impact of the virus[4], whether it will spread extensively in the heterosexual population in developed countries and why heterosexual transmission is so rapid in certain developing countries.

The potential for epidemic spread of HIV infection, and consequently AIDS, within a specific risk-group depends on the magnitude of the basic reproductive rate, R_0, for HIV within that group[5-8]. R_0 essentially measures the number of infections produced, on average, by an infected individual in the early stages of the epidemic, when virtually all contacts are susceptible. As such, it depends on the average probability, β, that an infected individual will infect a susceptible partner over the duration of their relationship, times the average number of new partners acquired per unit of time, c (within the specified risk category), times the average duration of infectiousness, D; $R_0 = \beta cD$. Setting aside problems associated with what exactly is meant by 'average' in the above statement, current data are not sufficient to assign accurate values to any one of these three components. In this article we describe recent progress toward such assignments, indicate some of the new problems and uncertainties that are emerging, and summarize such predictions as we believe can tentatively be made.

Incubation and infectious periods

Most epidemiological models of the spread of HIV assume some uniform level of infectiousness throughout the long and variable incubation period that elapses between HIV infection and the onset of AIDS[5-9]. Information about the distribution of incubation intervals mainly derives from transfusion-associated AIDS cases and cohort studies of patients for whom the time point of seroconversion is known. Most information derives from transfusion associated cases (these may be unrepresentative of incubation periods in patients infected by needle-sharing or sexual activity, or indeed for patients in developing countries[4,10,11]). A recent analysis[12] of 512 cases in which dates of transfusion and diagnosis of AIDS are known suggests the mean and median incubation periods are both between 7 and 8 years in patients older than 12 yr (based on a Weibull distribution for the incubation intervals where the probability to develop AIDS at time τ after infection, $p(\tau) = \exp(-a\tau^c)$ and a and c are constants). In past studies the estimated average incubation period has been disconcertingly close to the time span over which data are available, suggesting that the average could lengthen as more information accumulates. The most recent analysis, however, suggests that estimates of the average are settling around 7-8 yr. A similar estimate is reported for homosexual men in a cohort study in which dates of seroconversion were recorded[13]. The average is somewhat less in children (<12 yr) and elderly people (>60 yr), and appears to be much less in infants who acquired infection from their mothers[14].

Estimates of the fraction, f, of those infected with HIV who will go on to develop AIDS have increased with time. Rates of progression from seropositivity to AIDS (derived largely from cohorts where the dates of seronconversion are not known) range from <1% to >12% with an average of around 5-7% per annum[14-20]. Rates of progression to AIDS are much higher in perinatally infected infants[19]. Current evidence suggests 30-75% of infected individuals will have progressed to AIDS within 6 yr. In the San Francisco City Clinic cohort study of 6,700 homosexual and bisexual men, 75% had been infected with HIV within 88 months, and of these only 20% had remained completely asymptomatic[20]. Thus the possibility exists that f is essentially 1.0, with the only uncertainty being the distribution of times to the development of disease.

Once AIDS is diagnosed the mean life expectancy of patients reported in recent studies in the United States[21] and the United

Table 1 Probabilty of transmission

Risk group of index case	Probability of transmission	Sample size	Ref.
Male to female			
Transfusion	0.18	55	32
Mixed including bisexuals	0.23	97	33
Army personnel	0.33	18	34
Intravenous drug users	0.48	88	34
Mixed including bisexuals	0.50	28	36
Patients from Africa	0.73	38	37
Patients from Africa	0.71	14	46
Haemophiliac	0.03	30	38
Haemophiliac	0.04	77	39
Haemophiliac	0.07	148	40
Haemophiliac	0.10	21	41
Haemophiliac	0.10	40	47
Haemophiliac	0.17	21	42
Haemophiliac	0.06	33	43
Transfusion	0.16	50	44
Intravenous drug users	0.35	48	45
Female to male			
Transfusion	0.08	25	32
Army personnel	0.33	6	34
Patients from Africa	0.40	10	37
Intravenous drug users	0.58	12	35
Mixed	0.71	17	36
Male to male			
Male homosexual	0.60	35	48
Male homosexual	0.10	343	49

* Probability of transmission per sexual partnership expressed as a proportion of partners infected in the study sample.

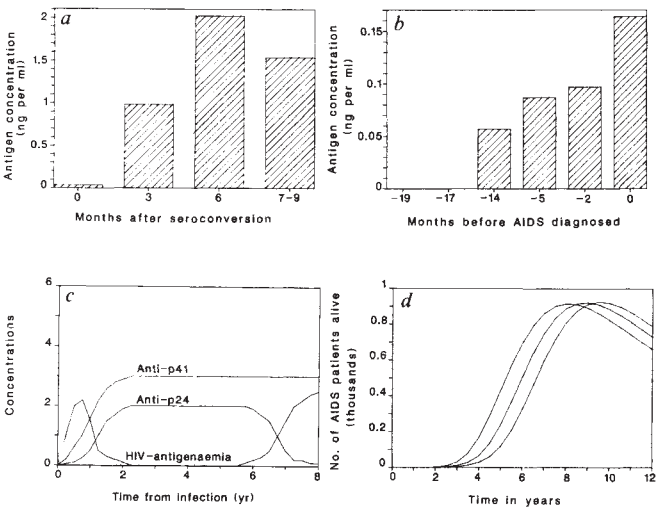


Fig. 1 *a*, Average concentration (ng per ml) of HIV antigens in serum samples taken from homosexual males infected with HIV at various times after seroconversion (data from refs 25 and 26). Samples sizes; $n = 11$ at time 0, $n = 8$ at 3 months, $n = 4$ at 6 months and $n = 4$ at 7–9 months. *b*, Antigen concentrations (ng per ml) at specific times (in months) before conversion to AIDS in serum samples drawn from a male homosexual patient (from refs 25, 26). *c*, A schematic representation of changes in antigen and antibody (p24 and p41) over the interval from infection to the development of AIDS (an average incubation period of 8 years)[24]. *d*, Numerical simulations generated by a simple model of two periods of infectiousness (transmission probabilities B_0 and B_1) split by a period of no infectivity[9,26,27]. The projections record the total number of AIDS patients alive at time t, where in the top curve at $t = 4$, the transmission coefficient for the first period of infectiousness (B_0) was set at 0.15 yr^{-1}, while the coefficient at the second episode (B_1) was set at 0.05 yr^{-1}, in the middle curve at $t = 4$, $B_0 = 0.05\ yr^{-1}$ and $B_1 = 0.15\ yr^{-1}$. The first and second episodes were assumed to last for an average 1 yr each, separated by an average period of 6 yrs of zero infectiousness. The epidemic was initiated in a population of 9,995 male homosexuals (homogeneously mixing) by the introduction of 5 infected persons. Recruitment of susceptibles was set at 200 yr^{-1}, life expectancy of an AIDS patient at 1 yr and the effective rate of partner change was set at 20 yr^{-1} ($R_0 = 4.0$). Note the magnitude of B_0 in the early episode is a major determinant of the rate of growth of the epidemic in its early stages[9,25,26].

Kingdom[22] is approximately a year. Clinical manifestation has the greatest influence on survival (those who presented with Kaposi's sarcoma had the most favourable survival rate), followed in importance by age, race or ethnicity, risk group and sex. Studies of survival in infants and children who acquired infection perinatally or by transfusion reveal shorter survival times, with a median of around 8 months[23].

Epidemiological and clinical studies suggest that the infectiousness of individuals with HIV may fluctuate over the long and variable incubation period between infection and AIDS. Longitudinal studies of fluctuations in viral abundance, or in the concentration of HIV antigens in serum, cells or semen, reveal great variability in temporal trends between patients, both in antigenaemia and antibody titres[24,25]. A tentative hypothesis is that antigenaemia rises to a peak during, or just before, primary HIV infection (Fig. 1*a*) and then typically falls to very low levels (often undetectable by current methods) before beginning to rise again as symptoms of disease appear and the persistent generalized lymphadenopathy progresses to the AIDS-related-complex and finally to AIDS (Fig. 1*b*). Patients who progress through the last step rapidly may have a detectable antigenaemia throughout the incubation period. Concomitant with these changes, antibodies (anti-p24) to core antigens rise slowly, reach peak titres when antigenaemia is undetectable or low, and then fall to very low levels as clinical disease develops (Fig. 1*c*). Antibodies to envelope antigens appear, on average, to remain high throughout the incubation period following primary HIV infection[24,25].

There are practical problems in the quantification of antigen titres in serum samples, the detection of infected cells and in relating antigen concentration in serum to infectivity (because it takes no account of virus abundance in secretions and excretions). However, if the two peaks in antigenaemia are associated with an early and late peak in infectivity, then simple mathematical models of the transmission dynamics of HIV indicate that the variability of infectivity over a long average incubation period can have significant implications for the interpretation of temporal trends in the incidence of AIDS[8,9,26,27]. In the simplest case the basic reproductive rate R_0 may be defined as $R_0 \sim c(\beta_0 T_0 + \beta_1 T_1)$, where c is the effective average rate of sexual partner-change and β_0 and β_1 are the transmission probabilities associated with the early and late episodes of infectivity.

That is to say new infections are produced at a rate $\beta_0 c$ over an average interval T_0 early in the course of infection; later infections are produced at a rate $\beta_1 c$ over an average interval T_1 as the AIDS-related-complex and then AIDS develop. Exponential growth in the early stages of the epidemic will tend to be set by the average transmission probability in the first episode of infectivity while the generation of secondary cases (R_0) will involve both episodes. In the terminology of conventional demography, early 'births' count more than later ones toward population growth rates, but all 'births' are relevant to the total number of offspring (here meaning infected people).

Two peaks in infectiousness will induce complex patterns of growth in the epidemic curve. As illustrated in Fig. 1*d*, an early rapid phase of growth in the incidence of AIDS will be followed by a longer period of slower exponential growth before saturation effects (stemming from exhaustion of the supply of susceptible) induce a decline in the incidence. At present, however, the validity of assuming two peaks in infectiousness remains in doubt due to the wide variability in antigenaemia observed in different patients through time. There are hints that burst of viraemia can occur between primary HIV infection and the development of disease; the simplest models that mimic the interaction between HIV and the immune system can generate such phenomena as a consequence of chaotic dynamical behaviour (R.M.A. and R.M.M., in preparation).

Probability of transmission

The main source of information about the probability of transmission is longitudinal studies of the likelihood of seroconversion for susceptible sexual partners of HIV-infected patients, such as monogamous partners of men or women infected either in the course of intravenous drug abuse or by transfusion of blood or blood products. The interpretation of such studies, however, is beset with many difficulties owing to the high variability in sexual practices and activity within small study populations, to problems in associating a transmission event with the length of time that the index case has harboured HIV (along with the time scale on which he or she will develop AIDS), and to the need to link the likelihood of transmission with highly variable quantitative measures of viral or antigen abundance, immunological parameters (CD4/CD8 cell ratio, CD4-cell depletion and antibody titres) and clinical measures of disease status. Recent reviews[28–31] suggest that the average transmission probability, β, for heterosexual partnerships is around 0.1–0.2 (uncertain to at least a factor 2) in studies with moderate sample sizes (Table 1). Surprisingly, according to a careful study by Peterman and colleagues, β is uncorrelated with the number of sexual acts (with one transfusion-infected individual infecting

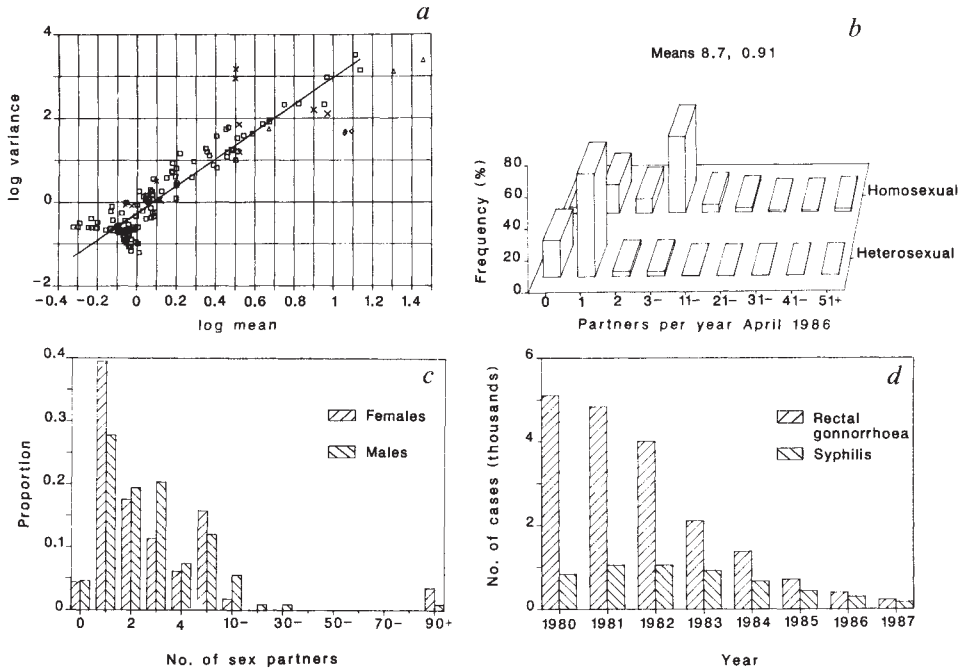


Fig. 2 *a*, A scatter plot of the logarithm (base 10) of the variance in reported numbers of different sexual partners versus the logarithm of the mean number of sexual partners from a wide range of published and unpublished studies[29]. The variances and means are drawn from studies that made use of a wide range of sampling methods, sampled different populations and recorded numbers of different sexual partners over differing time intervals (ranging from the past month to lifetime). The squares record data from surveys of heterosexual men and women in the general population of England and Wales, the crosses record data for heterosexual men and women attending sexually transmitted disease clinics (STD) in London, the triangles record data from male homosexuals in England and Wales and the diamonds record data for heterosexual males from a rural community in a Sub Saharan African country. The solid line denotes the best fit power model of the form $\sigma^2 = am^b$ with $a = 0.555$, $b = 3.231$, excluding the data from Africa. *b*, Frequency distributions of the number of different sex partners reported over a year period in samples (random location quota sampling) of male homosexuals and male and female heterosexuals in England and Wales. The data were collected in April 1986 and record number of different partners over the previous year (data from ref. 51; see ref. 29). The mean numbers of reported partners were 8.7 and 0.91 for the homosexual and heterosexual samples, respectively. *c*, Frequency distributions of the number of different sex partners over the previous year in samples of male and female heterosexuals attending a London STD clinic in March 1987 (data from ref. 29). *d*, The decline in the incidences of rectal gonorrhoea in males, and primary/secondary syphilis in homosexual males, in San Francisco, 1980–87 (data from ref. 52).

a spouse after one act, while others failed to do so after several hundred)[32]. Evidence about β among male homosexuals is more indirect, and comes mainly from using epidemiological models to analyze serological data that is combined with information about rates of acquiring new partners; the characteristic figure from studies with large sample sizes is again $\beta \sim 0.1$ for partnerships. The rough equivalence between estimates of β from male homosexual partnerships and for monogamous heterosexual partnerships may be because the transmission probability per act is higher for the former, but the average duration (and number of acts) per partnership is compensatingly higher for the latter. We are not aware of any estimates of β among drug users sharing needles.

Rates of acquiring sexual partners

Simple models of the transmission of HIV suggest that the 'effective average' rate of acquiring new sexual partners, c, should be interpreted as the mean rate, m, plus the variance to mean ratio σ^2/m, that is $c = m + \sigma^2/m$ (refs 5 and 7). Measurement of c therefore requires information on the full distribution of partner-change rates in a defined risk group. Recent surveys reveal great variability in these rates, not simply between individuals in a given community ($\sigma^2 \gg m$), but between different risk groups and different countries or communities[30,35]. The design and execution of surveys of sexual behaviour present many problems in sampling, assessment of reliability and interpretation[30]. Interestingly, however, a wide range of surveys of different populations employing different sampling methods and various time intervals for recall, reveal a remarkably consistent trend in the relationship between the mean and variance in the rate of acquisition of new partners (Fig. 2*a*). The summary statistics are related by a power law, $\sigma^2 = am^b$, where a and b are constants.

In general, the surveys reveal higher rates of partner change among male homosexuals than among heterosexuals, among attendees at sexually transmitted disease (STD) clinics (Fig. 2*c*) than among the general population, among teenagers and young adults than among older people, and among males in certain sub-Saharan African countries than among males in developed countries[8,30]. As might be expected, a history of past STD infection is often indicative of a high rate of partner change. The contrast between male homosexuals and heterosexuals in developed countries is often striking; a UK quota sample in April 1986, for example, suggested a factor 8 difference in the mean rates of sexual partner-change (Fig. 2*b*)[51]. However, a variety of surveys suggest that in recent times male homosexuals have greatly reduced their rates of partner-change (which is also reflected in trends in syphilis and rectal gonorrhoea in males, Fig. 2*d*) while little change is found among heterosexuals in developed countries. Despite the growing number of small scale surveys of sexual behaviour in specific communities, there remains a need for large scale national surveys (based on random sampling and on, say, a year by year basis) to assess the potential for the spread of HIV and the impact of education.

Serological surveys

The above discussion is organized around the components of R_0, whose magnitude determines whether an epidemic will take off in a particular risk group ($R_0 > 1$) or not ($R_0 < 1$). The detailed dynamics, however, depend on the probability distributions of the transmission probabilities, rates of partner change, and intervals of incubation and infectiousness. Information on HIV infection derived from serological surveys over time can be used to make indirect inferences about the transmission parameters, or to test theoretical predictions about how heterogeneities (for instance, in degrees of sexual activity) affect the shape of temporal trends in seroprevalence. An excellent recent review of such data in the United States by the Centres for Disease Control illustrates the value of drawing together published and unpublished seroprevalence data, and the need

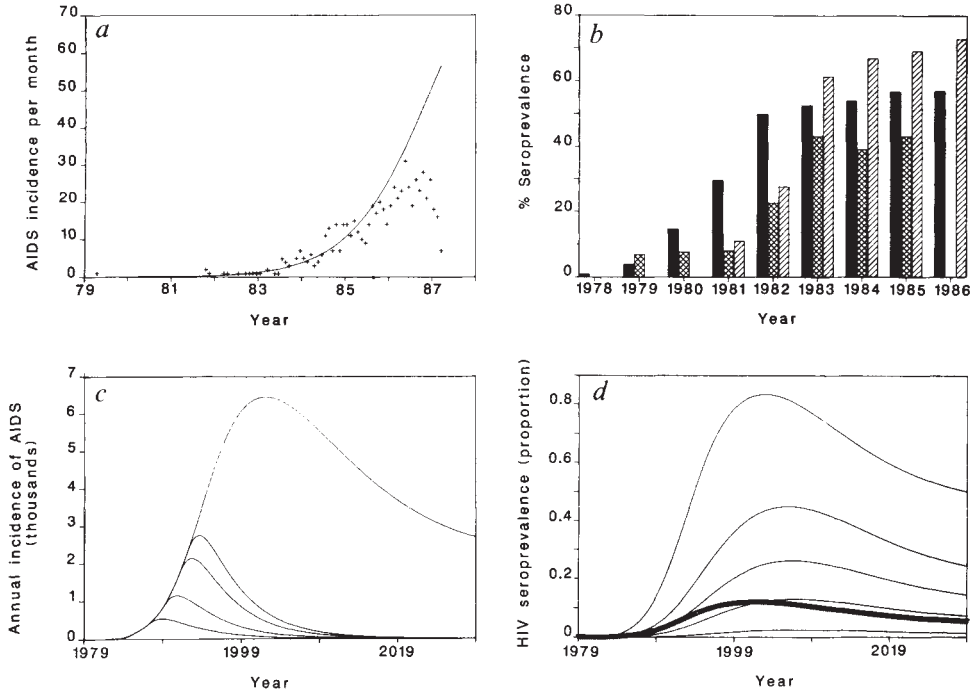


Fig. 3 *a*, Monthly incidence of AIDS as reported to the Public Health Laboratory Service in England and Wales up to the end of March 1987. The crosses denote the number of reports and the solid curve denotes the predicted number of cases once reporting delays (between diagnosis and report) are taken into account (D. R. Cox and G. F. Medley, personal communication). The short-term predictions are based on a logistic model. *b*, Changes in the percentage seropositive for antibodies specific to HIV antigens in longitudinal studies of male homosexual/bisexual men in San Francisco (a Cohort study, data from ref. 52, Table 12) (left-hand histogram bar) and intravenous drug users in Italy (data from ref. 7) (middle histogram bar, patients hospitalized for acute viral hepatitis in Milan; right-hand histogram bar, patients attending for melhadone maintenance treatment in Milan). Panels *c* and *d*, numerical simulations of the AIDS epidemic in the male homosexual/bisexual population of England and Wales. The projections are based on a hybrid model that incorporates distributed infectious and incubation periods (a mean of 8 years with an early phase of infectiousness lasting 6 months with $B_0 = 0.035$ and a late phase lasting 1.5 yrs with $B_1 = 0.07$) and heterogeneity in sexual activity (S. P. Blythe and R.M.A., in preparation). The mean and variance of the sexual activity distribution are defined by the power law relationship recorded in the legend to Fig. 2*a*. In *c*, the epidemic was initiated in 1979 by the introduction of one infectious male (first-episode) in a high sex activity class (20-50 partners per year), into an estimated population of 500,000 homosexual and bisexual men (4% of the sexually active male population between the ages of 16-56 years). In 1979 the mean rate of partner change was assumed to be 8 yr^{-1}. This average was assumed to decline to 6 in 1982, 4 in 1984 and 3.5 in 1987. The projected number of cases closely mirror observations over the period 1979 to 1987 in England and Wales. In the top line the mean rate of partner change was assumed to remain constant at 3.5 yr^{-1} for the remaining years of the simulation. From the bottom line to the second from the top in year 1999, transmission was halted (by the mean rate of partner change declining to 1.0 yr^{-1} such that $R_0 < 1$) in 1987, 1989, 1991 and 1992 respectively. Note how the projected incidence of AIDS depends critically on current patterns of sexual partner change and the large number of cases of AIDS that will appear in future years even after transmission of HIV is halted. In *d*, temporal changes in seroprevalence for HIV antibodies are recorded for the simulation (top line in *c*) in which the mean rate of sexual partner change declined to 3.5 yr^{-1} in 1987 and remained constant thereafter. From top to bottom at year 1999 the thin lines denote seroprevalence in groups of the population with sex partner change rates of 0-2, 2-5, 5-10, 10-20, 20-50 (yr^{-1}) respectively. The thick line denotes the overall seroprevalence in the total populations of homosexual/bisexual men.

for nationwide surveys[52]. Other countries should attempt similar exercises.

In the general population of the United States (assessed via surveys of blood donors, civilian applicants for military service, Job Corps entrants, sentinel hospital patients and women in settings related to fertility and childbearing), HIV infection prevalence is generally a fraction of 1%, although it varies widely in different locations and has been found to be much higher in selected inner city populations[18,53] (Fig. 3). Similar wide variation was found in high risk groups samples in different locations, such as male homosexual and bisexual men (<10%->70%, mostly in the range 20%-50%), intravenous drug abusers (<5%->60%), people with coagulation disorders (70% for haemophilia A and 35% for haemophilia B), heterosexual partners of infected people (<10%->60%) and female prostitutes (0.0%->40%)[52]. Such variability is generated by a variety of factors including the timing of the introduction of infection in an area and heterogeneity in sexual and other behaviours.

The most important factors in survey work are the assessment of changes through time and at different spatial locations; information of this kind is less readily available (Fig. 3). In US military service applicants and first-time blood donors, HIV prevalence appears to have remained stable. But great caution is needed in interpreting patterns of apparent stability because factors such as self-deferral of people with risk factors, balanced by new infections, plus a very slowly spreading heterosexual epidemic (perhaps with a doubling time of 6-14 yr, given observed differences in rates of sexual partner-change between male homosexuals and heterosexuals), offer alternative explanations to that which assumes the virus is not spreading in heterosexual communities.

Predictions and uncertainties

Fitting simple functions to incidence data and extrapolating beyond current observation can help in short-term prediction (one to a few years) and statistical models have been developed to take account of distributed delays between the diagnosis and reporting of a case to a central body (Fig. 3*a*). The average delay in Europe and the United States is increasing as the epidemic grows and this factor is largely responsible for the apparent decrease in the rate of exponential growth in the incidence of AIDS. Once reported delays are taken into account, the decrease is much less apparent. The incidence of new infections (seroconversions) in cohorts of male homosexuals has declined in recent years in developed countries but transmission continues unabated in other groups such as drug users (Fig. 3*b*).

Longer term prediction, based on (sometimes complex) models of viral transmission within and between different at-risk groups, is fraught with problems. An example of an attempt to project the course of the epidemic in the UK male homosexual

community, based on a model with variable incubation and infectious periods, and on heterogeneity in sexual activity, is given in Fig. 3*c*, *d* (ref. 55). The quantitative details of such projections are sensitive to assumptions concerning, for example, the fraction who developed AIDS (f), the average incubation period and the mean and variance of the rate of sexual partner-change (and how this changes through time). The timing of maximum incidence from the start of the epidemic is largely set by the average incubation/infectious period and the magnitude of R_0. The total epidemic is made up of a series of waves passing through the different sexual activity groups; those with high rates of partner-change are infected early on and those with low rates much later. Similarly, seroprevalence (% infected) is high in high partner change groups and low in low partner-change groups (Fig. 3). During the course of the epidemic the mean rate of sexual partner-change in the population will decline as a consequence of the early infection and death from AIDS of those individuals with high partner-change rates. This trend will occur irrespective of the impact of education on the sexual behaviour of individuals; it therefore complicates assessment.

A further complication is raised by the growing belief that HIV transmission in heterosexual and male homosexual communities can be facilitated by the presence of other STDs (so called co-factors)[56]. Such STDs are more frequently present among those with high rates of partner-change, further enhancing the disproportionate role of such individuals in acquiring and transmitting HIV infection. Simple transmission models of HIV infection in conjunction with another STD suggest that the control of HIV spread by treating the other STD is a sensible option provided the enhancing effect of the other STD on HIV transmission is strong[8]. This remains uncertain at present.

Transmission models certainly help in the interpretation of observed trends, but the uncertainties about key parameters, such as changes in sexual behaviour during the course of the epidemic, argue against placing too much reliance on long term predictions. The primary role of the development and analysis of transmission models at present is to focus attention on what needs to be measured to understand the future course of the epidemic. One thing that is certain, however, is that the epidemic will be long and drawn out as it spreads through the different at-risk groups and in different localities over the coming decades.

Heterosexual spread in developed countries

Growing evidence of HIV infection in heterosexual males and females with no known risk factors, other than being a regular sexual partner of an intravenous drug user or bisexual male (in the case of females), prompts the question of whether there will be a widely disseminated epidemic in heterosexual communities in developed countries[52,53,57]. It is not possible to answer this question with any certainty at present but consideration of what we do know about the components of R_0 provides some clues. Suppose the probability of transmission per partner contact, β, is 0.1 and the average duration of infectiousness over the long incubation period (~8 yr) is of the order of two years, then for R_0 to exceed unity in value the effective rate of sexual partner-change c (where $c = m + \sigma^2/m$, m is the mean and σ^2 the variance) must be of the order of five or more. As illustrated in Fig. 2*a* (data mainly from the United Kingdom) typical means and variances of partner change rates per annum in the population are around 0.5-1.5 and 1-5, respectively. Thus, for the general population, the value of R_0 may be just above or just below unity. If it is just above (say $\beta = 0.1$, $c = 5.5$, $D = 2$, giving $R_0 = 1.1$) then the doubling time of the epidemic during the early stages could be of the order of 8-14 years (a figure to be contrasted with the 1-year doubling time in the male homosexual population). It will therefore be extremely difficult to detect changes in seroprevalence in the general heterosexual population when, for example, the level of seroprevalence is a fraction of 1% (as it currently seems to be in the United States[52,53] and the United Kingdom[57,58]). The need for widescale serological screening is clear from this argument.

Our assignment of values to the components of R_0 for HIV spread in the heterosexual population is very crude, but one factor that is clear is the strong likelihood that the virus will spread in the high partner-change groups of the population. Attendees at STD clinics are particularly at risk given their above-average rates of sexual partner-change (Fig. 2*a*) and the liklihood that other STD infections increase the magnitude of β. The likely scenario for HIV spread is therefore a very slow rise in seropositivity in the general population on a time scale of decades, with infection largely restricted to the 'tail' of the sexual activity distribution. This may be too conservative a view if a proportion of those infected remain 'asymptomatic' but infectious over very long periods of time (say 10 years or more).

Conclusions

That many uncertainties still surround the key epidemiological parameters that determine transmission within and between different at risk groups is not surprising given the very long incubation period of AIDS and the primary mode of transmission of HIV. Epidemiological data are improving but there are important gaps in our understanding. These could be filled by more determined attempts, at national levels, to promote, adequately support and co-ordinate large scale, long term studies of changes in seroprevalence through time both in cohorts and the general population, of sexual lifestyles and behaviour (of particular importance in developing countries at present) and of the infectiousness of infected people. Such studies should include measurements of fluctuations in viral abundance in blood, secretions and excretions, along with immunological scores of disease progression and genetic changes in the virus population[59,60].

The need for a better epidemiological understanding is an essential prerequisite for the design of trials for promising candidate vaccines or drugs. It is also, in part, an insurance against failure to develop vaccines or effective drugs. The problems associated with testing a candidate vaccine are daunting, but would be made less so if more were understood about infectiousness and transmission. With respect to drugs, it remains to be proven that azidothymidine suppresses the infectiousness of treated patients and, unfortunately, any drug that prolongs survival could in principle enhance the transmission potential (by increasing D) of an infected person. The main hope for preventing a heterosexual wave of the epidemic in developed countries, given the failure to prevent extensive spread in the male homosexual and intravenous drug using populations, remains education and publicity. Without firm commitment and intensive effort, it is likely to be difficult to maintain public interest in the face of a very slowly developing heterosexual epidemic.

We gratefully acknowledge financial support from the Medical Research Council (R.M.A.) and the National Science Foundation (R.M.M.), and help and advice from A. Johnson, G. Medley, D. R. Cox, S. Blythe, S. Blower, E. Konings and J. Bull.

1. Centres for Disease Control *Morbid Mortal Wkly Rep.* **30**, 250-242; **30**, 350-308 (1981).
2. Barré-Sinoaussi, F. *et al. Science* **220**, 868-871 (1983).
3. Gallo, R. C. *et al. Science* **220**, 500-503 (1984).
4. Anderson, R. M., May, R. M. & McLean, A. R. *Nature* **332**, 228-233 (1988).
5. Anderson, R. M. *et al. IMA J. Math. appl. Med. Biol.* **3**, 229-263 (1986).
6. Anderson, R. M. & May, R. M. *Phil. Trans. R. Soc.* **B314**, 533-570 (1986).
7. May, R. M. & Anderson, R. M. *Nature* **326**, 137-142 (1987).
8. May, R. M. & Anderson, R. M. *Phil. Trans. R. Soc.* B (in the press).
9. Anderson, R. M. *J. R. Stat. Soc. Ser. C.* (in the press).
10. Medley, G. F. *et al. Nature* **328**, 718-721 (1987).
11. Medley, G. F. *et al. Proc. R. Soc.* **233**, 367-377 (1988).
12. Medley, G. F. *et al. Nature* (in the press).
13. Curran, J. W. *et al. Science* **239**, 610-616 (1988).
14. Rogers, M. F. *et al. Paediatrics* **79**, 1008-1014 (1987).
15. Jaffe, H. W. *et al. Ann. intern. Med.* **103**, 210-214 (1985).
16. Polk, B. F. *et al. New Engl. J. Med.* **316**, 61-66 (1987).
17. Moss, A. R. *et al. Br. med. J.* **296**, 755-750 (1988).

18. Titti, F. *et al. J. med. Virol.* **23**, 241-248 (1987).
19. Goedert, J. J. *et al. Science* **231**, 992-995 (1986).
20. Eyster, M. E. *et al. Ann. intern. Med.* **107**, 1-14 (1987).
21. Rothenberg, R. *New Engl J. Med.* **317**, 1297-1302 (1987).
22. Reeves, G. K. & Overton, S. E. *Lancet* **i**, 880 (1988).
23. Piot, P. *et al. Science* **239**, 573-579 (1988).
24. Pedersen, C. *et al. Br. med. J.* **295**, 567-569 (1987).
25. Goudsmit, J. & Paul, D. A. *Epidem. Inf.* **99**, 701-710 (1987).
26. May, R. M., Anderson, R. M. & Johnson, A. M. (in preparation).
27. Blythe, S. P. & Anderson, R. M. *IMA J. Math. appl. Med. Biol.* (in the press).
28. May, R. M. *Nature* **331**, 655-656 (1988).
29. Anderson, R. M. & Johnson, A. M. in *AIDS and Sex: An Integrated Biomedical and Behavioural Approach* (Kinsey Institute, in the press).
30. Blythe, S. P. & Anderson, R. M. *IMA J. Math. appl. Med. Biol.* **5**, 1-19 (1988).
31. Johnson, A. M. *J. Roy. Stat. Soc. Ser. C.* (in the press).
32. Peterman, T. A. *et al. J. Am. med. Ass.* **259**, 55-58 (1988).
33. Padian, N. *et al. J. Am. med. Ass.* **258**, 788-790 (1987).
34. Redfield, R. R. *et al.* III Int. Conf. on AIDS, Washington (1987).
35. Steigbigel, N. H. *et al.* III Int. Conf. on AIDS, Washington (1987).
36. Fischl, M. A. *et al. J. Am. med. Ass.* **257**, 640-644 (1987).
37. Taelman, H. *et al.* III Int. Conf. on AIDS, Washington (Abstr.) (1987).
38. Miller, E. J. *et al.* III Int. Conf. on AIDS, Washington (Abstr.) (1987).
39. Jones, P. *et al. Br. med. J.* **229**, 695-699 (1985).
40. Allain, J. P. *New Engl. J. Med.* **315**, 517 (1986).
41. Kreiss, J. K. *et al. Ann. intern. Med.* **102**, 623-626 (1985).
42. Winklestein, W., Wiley, J. A. & Padian, N. *J. Am. med. Ass.* **255**, 901-902 (1986).
43. Jason, J. M. *et al. J. Am. med. Ass.* **255**, 212-215 (1986).
44. Peterman, T. A. & Curran, J. W. *J. Am. med. Ass.* **256**, 2222-2226 (1986).
45. Saltzman, B. R. *et al.* II Int. Conf. on AIDS, Paris (Abstr.) (1986).
46. Sewankambo, N. K. *et al. AIDS* **1**, 113-116 (1987).
47. Bikerfield, G. *et al. Scan. J. infect. Dis.* **18**, 497-500 (1986).
48. Weller, I. V. D. *et al. J. med. Virol.* **22**, 91-98 (1987).
49. Grant, R. M. *et al. J. infec. Dis.* **156**, 189-193 (1987).
50. Johnson, A. M. *et al. Lancet* (in the press).
51. *AIDS Advertising Campaign. Report on Four Surveys During the First Year of Advertising 1986-87* (British Market Research Bureau, London, 1987).
52. Centres for Disease Control *A Review of Current Knowledge and Plans for Expansion of HIV Surveillance Activities* (30 November, 1987).
53. Quinn, T. C. *et al. N. Engl. J. Med.* **318**, 197-203 (1988).
54. Albert, J. *et al. AIDS* **2**, 107-111 (1988).
55. Blythe, S. P. & Anderson, R. M. *IMA J. Math. Med. Biol.* (in the press).
56. Greenblatt, R. M. *et al. AIDS* **2**, 47-50 (1988).
57. France, A. J. *et al. Br. med. J.* **296**, 526-529 (1988).
58. Evans, B. A. *et al. Br. med. J.* **296**, 473-475 (1988).
59. Srinwasan, A. *et al. Blood* **69**, 1766-1770 (1987).
60. Cheng-Mayer, C. *et al. Science* **240**, 80-82 (1988).

AIDS: MYSTERIES

Mysteries of HIV: challenges for therapy and prevention

Jay A. Levy

Department of Medicine and Cancer Research Institute, University of California School of Medicine, San Francisco, California 94143-0128, USA

A number of problems still surround infection by the human immunodeficiency virus and the pathogenesis of AIDS. Solutions to the problems would provide valuable information for the development of antiviral therapy and a vaccine.

DESPITE substantial progress in our understanding of the human immunodeficiency virus (HIV), a number of mysteries remain concerning the virus, its target cells and the responses of the infected host. If a rational approach to treatment and therapy of AIDS is to be successful, solutions to the mysteries must be found. From my perspective, the four major questions are: what are the viral and cellular determinants of infection, what mechanisms generate the different HIV strains, how does HIV cause disease, and what determines the course from infection to disease?

What governs viral tropism?

Recognition that the CD4 antigen on the surface of human helper T lymphocytes is a major cellular receptor for the HIV envelope protein gp120[1] led to the inference that endocytosis of the virus by cells expressing the CD4 molecule is an initial step in infection. But although mouse cells expressing human CD4 (as a result of experimental transfection) can bind HIV to the cell surface, they do not produce virus[2]. Moreover, HIV can infect cells lacking CD4, such as brain astrocyte cell lines[3] and human fibroblasts[4], and the virus is detected in endothelial[5] and epithelial[6] cells of seropositive individuals. Together, these observations strongly suggest that there are also mechanisms that do not involve CD4 by which HIV initially interacts with some cells.

Fusion of the target cell membrane with the HIV transmembrane envelope protein gp41 is one possible alternative means of virus entry. It may operate alone or in concert with gp120. In support of a fusion process, infection by HIV appears to be *p*H-independent[7]. Moreover, part of the sequence of gp41 is similar to a sequence coding for the fusogenic glycoproteins of paramyxoviruses[8], and antibodies to the gp41 protein neutralize HIV[9]. In addition, antibody-dependent enhancement of virus infection can mediate HIV infection of cells[10]. As previously described for dengue and other viruses[11], virus-antibody complexes permit HIV infection of macrophages and T cells, most likely via the complement and/or Fc receptor.

These findings should prompt re-evaluation of the likelihood that soluble CD4 will be of major value in the prevention of HIV infection of cells and encourage the further search for methods to control the first steps in virus infection. It could be worth directing attention to the second conserved region of the viral gp120, which is not involved in CD4 attachment but is essential for early stages of infection[12].

The establishment of HIV infection requires several processes that are influenced by both viral and cellular factors. Strain variations in specific viral genes may, for example, account for the fact that, compared with blood isolates of HIV, isolates from the central nervous system replicate better in macrophages—the main HIV-expressing cell type of the brain[5]—than in T lymphocytes[13]. The properties of brain isolates may be reflected in the clinical observation of some infected individuals who have neurological defects without signs of immune deficiency[14]. Strain variations in specific viral genes such as the heterogenous *orf-B* region (see below) may also account for the lack of replication of some HIV isolates despite their efficient attachment to human $CD4^+$ T cells[15].

In other studies, cellular factors appear to determine virus replication. For instance, a single HIV isolate displays different levels of productive infection in peripheral blood mononuclear cells from various individuals[15]. Cultured mouse cells, in contrast to human cells, do not produce a substantial number of virus progeny after transfection with a biologically active DNA clone of HIV[16]. Cellular factors, such as the NF-κB protein, have been shown to interact with regulatory regions of HIV and enhance virus replication[17].

Defining the viral and cellular genes governing the host-range specificity of HIV is a major avenue of study. Experiments using

recombinant HIVs, in which portions of macrophage-tropic and lymphocyte-tropic strains are combined, or using site-directed mutagenesis of biologically active DNA clones of these viruses should help clarify the viral genes involved. Assays of proteins found in selected cells or the use of somatic cell hybrids (mouse-human, for example) could uncover cellular factors influencing viral tropism. Clearly, the wide host range of HIV and the factors affecting virus replication are important variables to be defined for therapeutic strategies against viral infection of all cell types.

What causes HIV heterogeneity?

Biological heterogeneity of HIV is reflected by differences in host range, replicative properties and cytopathic effects in infected cells[18]. Restriction enzyme[19] and sequence analyses[20] of HIV isolates, as well as the patterns of neutralization of different isolates by antibodies,[21] also demonstrate genetic diversity, particularly in the envelope glycoprotein. (This is true not only of HIV-1 but also of the more recently identified HIV-2 subtype[22]). The mechanism responsible for generating these varying strains of virions is puzzling. One theoretical possibility is that the unintegrated proviral copies of HIV that accumulate during acute replicative infection[23] can undergo efficient genomic recombination leading to the evolution of infectious variants. The lack of genetic changes in HIV during long-term passage of the integrated virus in persistently infected cells[24] is consistent with this hypothesis.

Three explanations may account for the occurrence of envelope variants. First, the immunological reaction of the host selects specific spontaneous mutants which differ in their external envelope region. Thus far, this mechanism has only been observed with certain animal lentiviruses[25]. Second, mutations in the envelope, in contrast to other structural genes, can be tolerated during HIV replication. Third—and most speculatively—the HIV reverse transcriptase may be most error-prone when dealing with sequences for glycosylated proteins.

What determines pathogenesis?

An understanding of how HIV causes disease is a major research objective. As with any infectious agent, both viral and host determinants are involved.

Neuropathy. The disease in the central nervous system has been attributed by some investigators to the production by infected macrophages of cytokines that affect normal brain function[5,26]. HIV-infected brain capillary endothelial cells could compromise the maintenance of the blood-brain barrier, so permitting entry of toxic materials from the blood[5,26]. HIV also infects oligodendrocytes and astrocytes[3,5,14]. Compared with brain macrophages, only a small number of these glial cells produce viral RNA[4,5] but persistent or low replication of HIV in glial cells might affect their function. Since astrocytes maintain the integrity of the blood-brain barrier[27] and oligodendrocytes produce myelin, which is required for nerve conduction, the participation of these infected cells in AIDS neuropathy should be further evaluated. Finally, whether the neurologic disease results from long-term effects of HIV infection or specific pathogenic properties of a neurotropic strain[13,14] remains to be elucidated.

Enteropathy. HIV infection of enterochromaffin cells in the intestinal mucosa[6], perhaps by a process similar to that in the brain, could explain the chronic diarrhoea and malabsorption (particularly with duodenal involvement) observed in AIDS patients in the absence of any known bowel pathogens. These neuroendocrine cells migrate during embryogenesis from the neural crest and help regulate motility and digestive functions of the intestine. How their infection specifically accounts for the pathology is not known and the possible toxic effects of cytokines secreted by infected macrophages in the lamina propria needs to be considered.

Immunopathology. How HIV causes immune suppression presents the greatest enigma. The immune dysfunctions identified in HIV infection were first linked to the cytopathic effects of HIV on helper T lymphocytes that play a vital role in regulating immune function. These cells infected with HIV *in vitro* undergo syncytial formation by recruiting many uninfected cells before proceeding to cell death[1].

Table 1 Immune function abnormalities in AIDS

T lymphocytes
(1) Decreased proliferative responses to mitogens, soluble antigens, and allogeneic cells.
(2) Decreased lymphokine production (IL-2, gamma interferon) in response to antigen.
(3) Decreased cytotoxic T lymphocyte activity against virus-infected cells.

B lymphocytes
(1) Polyclonal activation with hypergammaglobulinaemia and spontaneous plaque forming cells.
(2) Decreased humoral response to immunization.
(3) Production of autoantibodies.

Monocytes
(1) Decreased chemotaxis.
(2) Decreased IL-1 production (or production of an inhibitor of IL-1).
(3) Decreased microbiocidal activity.

NK cells
(1) Decreased cytotoxic activity.

These abnormalities appear to begin with acute depletion of T helper cells and proliferation of B cells; other defects, many revealed by *in vitro* studies, accumulate over time. Several of the immune abnormalities may result from the decrease in $CD4^+$ lymphocytes and cytokine production[32].

Several observations, however, have challenged this suggested explanation for HIV-induced immunopathology. First, despite extensive histopathological studies of infected individuals, little evidence exists for syncytial cell formation *in vivo*[28]. Second, the number of HIV-infected cells in the blood (10^2–10^3 per ml)[29] does not account for the quantity of cells lost over time in the infected host; moreover, $CD4^+$ cells infected by HIV can survive several weeks in culture[30]. Third, many HIV isolates obtained from immunologically suppressed individuals are not highly cytopathic *in vitro*[18,31] and non-cytopathic isolates of HIV have sometimes been associated with disease[22]. Finally, abnormalities of immune function are observed not only in HIV-infected helper T cells and macrophages but also in uninfected cells of the haematopoietic system (Table 1)[32].

What other mechanisms could explain the immunological features of HIV pathogenesis (Table 2)? HIV infection is known to cause epiphenomena including decreased expression of immunological recognition sites, such as CD4 and the interleukin-2 (IL-2) receptor, and reduced production of cytokines such as IL-1, IL-2, and gamma interferon (Table 2)[32]. Loss of these molecules could have far reaching effects on other cells of the immune system. Moreover, HIV infection of progenitor cells could reduce the replenishment of circulating lymphocytes and macrophages. The disarray in immune function could generate T suppressor cells or factors[33] that further compromise the immune response. HIV envelope proteins have been found to be toxic or inhibitory to immune cells[34]. By circulating in the blood, these proteins may block the ability of lymphocytes to recognize or respond to foreign antigens and may interfere with the function of antigen-presenting cells.

In addition, HIV is associated, often early in the infection, with a polyclonal activation of B cells resulting in hypergammaglobulinaemia[32]. Some of these antibodies form immune complexes that can be detrimental to the immune system; some react with self-proteins, leading to autoimmunity[35]. The destruction of activated helper T cells by lymphocytotoxic autoantibodies directed against a normal cellular protein (p18) has been suggested as one mechanism for the substantial loss of $CD4^+$ cells in advanced disease[35]. Antibodies and/or cytotoxic cells directed

against other shared antigens on immune cells can induce further abnormalities in immune function or haematopoiesis[36,37]. Finally, antibody-dependent cellular cytotoxicity reacting against envelope proteins bound to ligands on uninfected $CD4^+$ cells could be a factor in the pathogenic process[38]. Clearly, multiple effects of HIV infection on the immune system, both direct and indirect, need to be fully evaluated to appreciate their potential role in the immunological abnormalities observed.

Cytopathology. A critical question surrounding HIV pathogenesis is how the virus kills the cell. In general, helper T lymphocytes are most susceptible to the cytopathic effects of HIV; in culture, they often undergo fusion before cell death. Macrophage killing by some strains of the virus has also been observed[1]. Syncytial formation by T cells, however, does not always preceed cell death; infected cells can die without undergoing fusion[39]. The viral envelope gp120, alone or through its interaction with CD4, could be responsible for cell death[34,40,43]. Direct inoculation of inactivated virus or the envelope gp120 onto peripheral blood mononuclear cells can also produce cytopathic effects[34,41]. Cell death may result from direct membrane disruption involving calcium channels[42] and/or phospholipid synthesis[43]. The accumulation of unintegrated proviral copies of HIV DNA is an attractive explanation for cytopathology since it is associated with cell death in other retrovirus systems[44]. Finally, whether the cytopathology of HIV in cell culture mirrors pathogenesis in the host awaits development of an appropriate animal model.

What influences progression to disease?

Long-term follow-up studies of HIV seropositive individuals indicate that about a third will remain free of symptoms for at least seven years[45]. Moreover, some healthy seropositive individuals lose a large proportion of their $CD4^+$ lymphocytes, yet do not develop symptoms[46]. Several AIDS patients (many with Kaposi's sarcoma) have remained clinically stable for up to six years. Two patients followed at our medical center have had only 5% of their normal $CD4^+$ cell number for over a year without any new symptoms. Thus, predictions on the development of opportunistic infections or cancers based solely on a decrease in $CD4^+$ cells could be misleading; other presently unrecognized functions of the immune system may be fundamental in warding off disease.

$CD8^+$ cell activity. What determines the resistant state? Cytotoxic T lymphocytes that react with cells expressing HIV proteins[47,48] have been noted in infected individuals, but the clinical importance of this antiviral response is unknown. In our laboratory, studies of asymptomatic individuals have revealed that their peripheral blood mononuclear cells do not readily release HIV when placed in culture. Nearly all, however, yield virus when a subset of their $CD8^+$ lymphocytes is removed from the blood sample[49]. Similar observations have recently been made with the primate immune deficiency virus[50]. These $CD8^+$ cells apparently prevent HIV replication not by killing infected cells, but by producing a diffusible suppressor factor or factors. Among seropositive individuals, the level of this $CD8^+$ cell activity varies, and can be reflected in clinical status. Peripheral blood mononuclear cells cultured from many patients with disease readily produce virus and their $CD8^+$ cells show very little antiviral activity.

This variation in HIV replication in cultured cells probably mirrors the increase of viral p25 antigen in the plasma of individuals as they advance in disease[51]. Whether this observation reflects enhanced virus production or a decrease in antibodies to p25 is still not clear. Nevertheless, the information does suggest that the resistant state in infected individuals is mediated by cellular immune responses operating soon after infection; once these are reduced, renewed production of HIV can occur. The resumption of HIV replication could enhance progression to disease because of the emergence, by mutation or selection, of new HIV variants that replicate rapidly to high titre in a variety of cell types, and that are highly cytopathic[18]. This observation correlates clinically with the increased loss of $CD4^+$ cells[46] and the high levels of p25 antigenaemia[51] associated with development of disease. Nevertheless, whether the progression to disease results first from a reduced immune response or from the eventual emergence of a more pathogenic HIV strain, or from both, needs to be clarified.

Taken together, the data suggest steps in HIV infection that might explain variations in the course of the disease (Fig. 1). Levels of antiviral immune response, the inherent sensitivity of the host cell to virus replication and the relative virulence of the virus strain are major variables to be defined. What ends the resistant state (that is, triggers progression to disease) is not known, but could be a variety of factors including antilymphocyte antibodies, enhanced production of virus by activating events, progressive destruction of $CD4^+$ lymphocytes by inter-

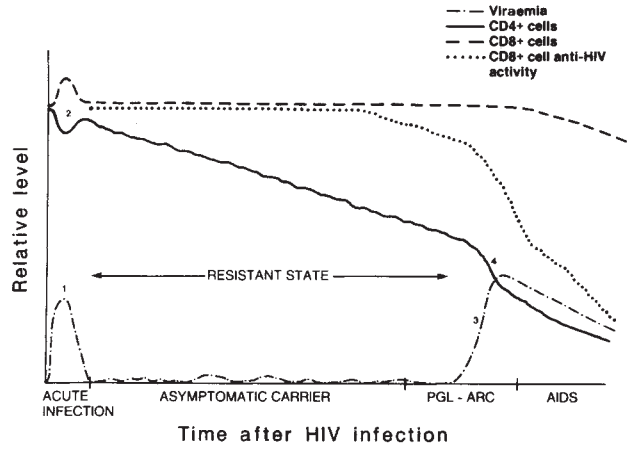


Fig. 1 A model of possible stages in the course of HIV infection. (1) Acute infection is characterized by the presence of free-virus (or its antigen) in the blood in the absence of antibodies to HIV. (2) A $CD8^+$ cells increase rapidly with a concomitant decrease in $CD4^+$ cells; thus the helper/suppressor ratio is dramatically reduced. After a short period (2–12 weeks) a resistant state develops in which $CD8^+$ cells return to above normal levels, $CD4^+$ cells decrease at a slow rate, and free virus is not readily detected in the blood, probably because it is produced episodically. Subsequently, after a variable time lag, the resistant state wanes for reasons that are not known. (3) The virus enters multiple replicative cycles during which a variant strain that can be resistant to the immune response and highly cytopathic can emerge and further compromise the antiviral state. (4) Viraemia is accompanied by enhanced destruction of $CD4^+$ cells and progression of disease.

Table 2 Mechanisms of immune suppression by HIV infection

Direct mechanisms
(1) HIV cytocidal effect on $CD4^+$ lymphocytes.
(2) Functional defects in infected $CD4^+$ cells:
  (*a*) decreased expression of cell surface proteins (for example, IL-2 receptor, CD4);
  (*b*) impaired production of lymphokines such as IL-2 or gamma interferon.
(3) Impaired antigen presentation and/or monokine production by infected macrophages; cell death.

Indirect mechanisms
(1) Generation of suppressor T cells and/or factors.
(2) Toxic or inhibitory effects of viral protein
(3) Immune complex formation.
(4) Induction of autoimmune phenomena:
  (*a*) autoantibodies resulting from polyclonal B cell activation or antigen mimicry;
  (*b*) virus mediated, enhanced immunogenicity of normal cellular proteins.
(5) Cytotoxic cell activity against viral or self proteins.

mittent periods of HIV replication and a decreased production by $CD4^+$ lymphocytes of cytokines required for the growth and function of antiviral $CD8^+$ cells.

Cofactors in pathogenesis. A major unresolved issue is whether other infectious agents can affect this asymptomatic period. Concomitant viral or parasitic infections may activate immune cells so they produce more virus, or become more sensitive to HIV infection[1]. Such events might lead to the suggested intermittent periods of HIV production. As demonstrated *in vitro*, agents such as the herpes viruses may act on the regulatory regions of HIV to enhance virus replication within the cell[52]. Continued surveillance of individuals with long asymptomatic periods, as well as studies of infected chimpanzees that have not yet shown any clinical abnormalities, should provide insight into factors influencing resistance to HIV.

Latency. A long incubation period might also be explained by a latent infection. During the latent period of retrovirus infection very little viral protein or RNA is made and no infectious progeny are produced by the infected cell[53]. When cells that have been latently infected with HIV in culture are treated with activating agents, such as halogenated pyrimidines and cytokines, virus replication begins and is then either maintained or reverts to latency[14,54].

Studies of the *orf-B* (or *3′ orf*) gene of HIV suggest it could be responsible for latency. Deletion of *orf-B* leads to a 5–10 fold increase in virus replication compared with the wild-type virus[55]. Conceivably, the *orf-B* gene product (p27 protein), which has GTPase and phosphorylating properties[56], interacts with cellular factors to down-regulate virus replication (see above) in a continuum that can proceed to latency.

In rare cases, individuals who have been HIV seropositive become seronegative. In some of these individuals the presence of a latent HIV infection in peripheral mononuclear cells can be detected by means of the polymerase chain reaction; in others no HIV can be detected[57]. This potentially encouraging observation suggests that HIV infection in some individuals might be eliminated completely, but most likely the virus remains latent at other sites. Defining the factors governing latency should provide valuable information for the development of antiviral strategies. Moreover, the importance of latency to viral transmission must be assessed in 'false negative' serological states.

Conclusions

Answers to many questions about the viral and host determinants of HIV pathogenesis could assist the prospects for therapy and prevention of AIDS. On current information, there are several features of HIV that need to be taken into account (Table 3).

Table 3 Features of HIV of relevance to antiviral therapy

(1) Virus infection involves integration of the viral genome into the chromosome of the infected cell. This cell is a protective environment for the virus and a reservoir for persistent virus production.
(2) The infected cell is a major source of virus transmission and can pass HIV by cell-to-cell contact.
(3) The infected cells can remain "latent" and express very few viral antigens. Can these cells be recognized and eliminated by an antiviral response?
(4) HIV transmission occurs at specific sites in the host (such as the rectum). Prevention requires immune response at these local sites.
(5) Several independent serotypes and subtypes of HIV can be identified. Can they all be controlled by one strategy?
(6) Portions of HIV proteins resemble normal cellular proteins. Immunization may induce autoantibodies.
(7) Vaccination may induce antibodies that enhance HIV infection.

We should target antiviral drug strategies at the vulnerable sites of HIV replication, both before and after integration. In this regard, understanding how an HIV strain evolves into a more cytopathic (and potentially pathogenic) agent[18] would be valuable. We must find methods for inducing strong intracellular and cellular host responses against the virus: intracellular production of the *orf-B* protein or stimulation of the $CD8^+$ cell population responsible for suppresssing HIV replication might produce long asymptomatic periods. Preventing the formation of antibodies to lymphocytes would be another promising direction. For vaccine development we need novel approaches that will define both specific epitopes of HIV and the appropriate adjuvant to elicit strong cross-reacting immune responses not generally observed in natural infection. Toward this objective, elimination of those epitopes responsible for antibody-dependent enhancement[10] would appear important. The immunized host must respond not only against free virus, but most importantly against productively and latently infected cells that can be major sources of HIV transmission[58]. Concentration on these areas of research should provide valuable information to help in the attack against HIV. In the process, we will learn a great deal more about viruses and the function of the immune system.

Studies conducted by the author were supported by the California State Universitywide Task Force on AIDS, the American Foundation for AIDS Research, and the National Institutes of Health. I would like to thank Drs S. Levy and J. Ziegler as well as Drs C. Cheng-Mayer, L. Evans, J. Homsy, J. Hoxie, J. Leong, M. McGrath and C. Walker for their helpful comments on this article, and C. Beglinger for its preparation.

1. Fauci, A. S. *Science* **239**, 617–622 (1988).
2. Maddon, P. J. *et al. Cell* **47**, 333–348 (1986).
3. Cheng-Mayer, C. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 3526–3530 (1987).
4. Tateno, M. & Levy, J. A. IV Int. Conf. AIDS, Stockholm (abstr.), (1988).
5. Wiley, C. A. *et al. Proc. natn. Acad. Sci. U.S.A.* **88**, 7089–7093 (1986).
6. Nelson, J. A. *et al. Lancet* **i**, 259–262 (1988).
7. Stein, B. S. *et al. Cell* **49**, 659–668 (1987).
8. Gallaher, W. R. *Cell* **50**, 327–328 (1987).
9. Chanh, T. C. *et al. EMBO J.* **5**, 3065–3071 (1986).
10. Robinson, W. E. *et al. Lancet* **i**, 790–795 (1988).
11. Halstead, S. B. & O'Rourke, E. J. *J. exp. Med.* **146**, 201–217 (1977).
12. Ho, D. D. *et al. Science* **239**, 1021–1023 (1988).
13. Gartner, S *et al. Science* **233**, 215–219 (1986).
14. Levy, J. A. *et al. Ann. Inst. Pasteur* **138**, 101–111 (1987).
15. Evans, L. A. *et al. J. Immun.* **138**, 3415–3418 (1987).
16. Levy J. A. *et al. Science* **232**, 998–1001 (1986).
17. Nabel, G & Baltimore, D. *Nature* **326**, 711–713 (1987).
18. Cheng-Mayer, C. *et al. Science* **240**, 80–82 (1988).
19. Hahn, B. H. *et al. Science* **232**, 1548–1553 (1986).
20. Starcich, B. R. *et al. Cell* **45**, 637–648 (1986).
21. Weiss, R. A. *et al. Nature* **324**, 572–575 (1986).
22. Evans, L. A. *et al. Science* **240**, 1522–1525 (1988).
23. Luciw, P. A. *et al. Nature* **312**, 760–763 (1984).
24. Robert-Guroff, M. *et al. J. Immun.* **137**, 3306–3309 (1986).
25. Carpenter, S. *et al. J. Virol.* **61**, 3783–3789 (1987).
26. Price, R. W. *et al. Science* **239**, 586–592 (1988).
27. Fontana, A. *et al. Nature* **307**, 273–276 (1984).
28. Cohen, M. B. & Beckstead, J. in *AIDS: Pathogenesis and Treatment* (ed. Levy, J. A.) (Dekker, New York, in the press).
29. Harper, M. E. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 772–776 (1986).
30. Hoxie, J. A. *et al. Science* **229**, 1400–1402 (1985).
31. Asjo, B *et al. Lancet* **2**, 660–662 (1986).
32. Koenig, S. & Fauci, A. S. in *AIDS: Etiology, Diagnosis, Treatment and Prevention*, 2nd edn. (eds DeVita, V., Hellman, S. & Rosenberg, S.) (Lippincott, Philadelphia, in the press).
33. Laurence, J. *et al. J. clin. Invest.* **72**, 2072–2081 (1983).
34. Shalaby, M. R. *et al. Cell Immun.* **110**, 140–148 (1987).
35. Stricker, R. B. *et al. Nature* **327**, 710–713 (1987).
36. Ziegler, J. & Stites, D. P. *Clin. Immunol. Immunopathol.* **41**, 305–313 (1986).
37. Donahue, R. E. *et al. Nature* **326**, 200–203 (1987).
38. Weinhold, K. J. *et al. Lancet* **i**, 902–904 (1988).
39. Somasundaran, M. & Robinson, H. L. *J. Virol.* **61**, 3114–3119 (1987).
40. Hoxie, J. *et al. Science* **234**, 1123–1127 (1986).
41. Rasheed, S *et al. Virology* **154**, 395–400 (1986).
42. Gupta, S. & Vayuvegula, B. *J. clin. Immun.* **7**, 486 (1987).
43. Lynn, W. S. *et al. Virology* **163**, 43–51 (1988).
44. Keshet, E. & Temin, H. M. *J. Virol.* **31**, 376–388 (1979).
45. Rutherford, G. W. & Werdegar, D. in *AIDS: Pathogenesis and Treatment* (ed. Levy, J. A.) (Dekker, New York, in the press).
46. Lang, W. *et al.* Abstr. Int. on Conf. AIDS, Stockholm (abstr.) (1988).
47. Walker, B. *et al. Nature* **328**, 345–348 (1987).
48. Plata, F. *et al. Nature* **328**, 348–351 (1987).
49. Walker, C. M. *et al. Science* **234**, 1563–1566 (1986).
50. Kannagi, M. *et al. J. Immun.* **140**, 2237–2242 (1988)).
51. Lange, J. M. A. *et al. Br. med. J.* **293**, 1459–1462 (1986).
52. Mosca, J. D. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 7408–7412 (1987).
53. Rojko, J. L. *et al. Nature* **298**, 385–388 (1982).
54. Folks, T. M. *et al. Science* **231**, 600–602 (1986).
55. Luciw, P. A. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 1434–1438 (1987).
56. Guy, B. *et al. Nature* **330**, 266–269 (1987).
57. Farzadegan, H. *et al. Ann. int. Med.* **108**, 785–790 (1988).
58. Levy, J. A. *J. Am. med. Ass.* **259**, 3037–3038 (1988).

Tidal disruption of stars by black holes of 10^6–10^8 solar masses in nearby galaxies

Martin J. Rees

Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK

Stars in galactic nuclei can be captured or tidally disrupted by a central black hole. Some debris would be ejected at high speed; the remainder would be swallowed by the hole, causing a bright flare lasting at most a few years. Such phenomena are compatible with the presence of 10^6–10^8 $M_\odot$ holes in the nuclei of many nearby galaxies. Stellar disruption may have interesting consequences in our own Galactic Centre if a $\sim 10^6$ $M_\odot$ hole lurks there.

TIDAL capture and disruption of stars was explored in the 1970s as a possible way of fuelling active galactic nuclei (AGNs). The outcome of these studies was that tidal disruption could not supply a quasar-level fuelling rate unless the stellar concentrations around the hole were so great that other processes (such as stellar collisions) would supervene and release still more gas[1-4]. The process has now, however, become interesting in an obverse context. The dynamics of stars in the inner regions of nearby galaxies such as M31 and M32 indicate the presence of central concentrated dark masses, and one would like to know whether these might indeed be massive black holes[5-9]. As the number of surrounding stars, and their motions, are roughly known, the capture rate by the putative central hole can be estimated; the question then arises of whether the apparent quiescence of the nuclei of these galaxies is compatible with a massive hole's presence. The answer depends on what happens to the debris from each disrupted star. How much is accreted or expelled? What is the associated luminosity or radiative efficiency? And how long does it take to digest or expel the debris from one star, in comparison with the interval between successive captures?

The conjecture that there might be a massive black hole in our own Galactic Centre goes back more than 15 years (see, for example, ref. 10). There is some dynamical evidence for a mass concentration within the central 0.1 pc (ref. 11); the peculiarities of the unresolved radio source[12] indicate some unique object right at the centre, even if much of the activity and structure outside the central parsec can be accounted for in conventional terms. But the situation here is still somewhat ambiguous: in a recent review, Genzel and Townes[13] rate the evidence for a black hole of order 10^6 $M_\odot$ as 'substantial but not fully convincing'.

Studies of the quasar and AGN population at redshifts $z > 2$, lead to the well known inference that dead remnants—black holes lurking in the centres of relatively quiescent galaxies—should be widespread. The lifetime of AGNs is uncertain: there could be one generation, with typical individual lifetimes $\gtrsim 2 \times 10^9$ years; alternatively, many successive generations could trace out relatively brief life cycles over the period when the entire population evolves. In the former case, quasars should leave rare and hugely massive remnants; in the latter, remnants would be more common, but less individually massive. In the latter case (but not the former), Seyfert galaxies could be interpreted as reactivated quasars. Evidence on the masses, and frequency of occurrence, of black holes in present-day galaxies, therefore offers important clues to quasar lifetimes, and to the relationships between different classes of AGNs; it may also indicate how many galactic mergers, leading to coalescence of black holes, have occurred since the epoch $z = 2$.

The best diagnostic for a black hole's presence would be some inevitable concomitant that cannot be explained in any other way. A candidate may be the distinctive manifestations of tidal capture and disruption of stars. If the stellar density in the central few parsecs is known, it is in principle straightforward to estimate how often a star gets close enough to the central hole to be tidally disrupted (or at least captured). The debris from one solar-mass star per ten thousand years, swallowed steadily with 10 per cent radiative efficiency, would yield a luminosity of 6×10^{41} erg s^{-1}—higher than is observed from M31 or M32, or indeed from our own Galactic Centre. Some authors, particularly Ozernoi and his collaborators[14-16], believe such considerations may actually rule out a 10^6 $M_\odot$ hole at our Galactic Centre.

Several features of the stellar capture process have as yet not been fully analysed, and create uncertainties in this line of argument: (1) What fraction of the debris goes down the hole, rather than being expelled? (2) What is the radiative efficiency for the accretion process? In other words, how many ergs are radiated for each gram that is swallowed? (3) How long does it take to 'digest' or expel the debris from one star? In particular, how does the flare duration and decay timescale for such a process compare with the interval between one stellar disruption and the next? Were this time much shorter than the mean interval between successive captures, one might expect bright flares with short duty cycles; observational constraints would not then be stringent until we had observed enough candidate holes to constitute a proper 'ensemble average'.

This paper is concerned with such phenomena, with particular reference to the nearest galaxies, where the masses of the holes (if they are indeed present) are perhaps of order 10^7 $M_\odot$. I consider the fate of stars that approach the hole closely enough to get captured, showing that some debris would be expelled at $\gtrsim 10^4$ km s^{-1}; the remainder accretes on to the hole, yielding high luminosities and perhaps a further phase of radiation-driven outflow. Some implications of putative black holes in our Galactic Centre and in other nearby galaxies are then described.

Gravitational influence on surrounding stars

There have been several analyses of how a central mass influences the orbits and spatial distribution of stars in the inner regions of its host galaxy[17-20]. A central mass M_h would influence the dynamics at all radii where $(GM_h/r)^{1/2} \gtrsim \sigma(r)$, the stellar velocity dispersion at r. If the stellar distribution in the galaxy had a well defined core, characterized by a radius r_C and a stellar density N_* (the virial velocity of the stars then being $\sigma \simeq (Gm_*/r_C)^{1/2}$), and the central hole had a mass $M_h \ll N_* r_C^3$, then a typical star in the core would spend most of its time moving under the combined influence of the other stars rather than the hole; the hole would dominate the dynamics only within a radius $r_h = GM_h/\sigma^2 = (M_h/N_* r_C^3) r_C$. In the case of M31, for instance

(refs 7-9), there is no well defined r_C—there is nonetheless evidence for a characteristic radius r_h within which σ rises (as does the rotational velocity) and the inferred mass-to-light ratio becomes anomalously large for a purely stellar population.

A star interacting with a massive hole cannot be treated as a 'point mass' if its gets so close to the hole that it becomes vulnerable to tidal distortions and even disruption. Such effects become important when the pericentre r_{min} becomes as small as the 'tidal radius'

$$r_T \simeq 5\times10^{12}\, M_6^{1/3}\, (r_*/r_\odot)\, (m_*/m_\odot)^{-1/3}\ \mathrm{cm} \qquad (1)$$

where M_6 denotes the hole's mass in units of $10^6\, M_\odot$. Essentially, r_T (which is $\ll r_h$) is the distance from the hole at which M_h/r^3 equals the mean internal density of the passing star. (r_T is, in order of magnitude, the same as the Roche radius: the latter, though a precisely defined quantity, is only strictly applicable to a star in a circular orbit with synchronized spin, which is not the relevant situation here.) The escape velocity from the half-mass radius of a solar-type star is $\sim10^3\ \mathrm{km\,s^{-1}}$, and its specific gravitational self-binding energy is $\sim10^{-5}\ c^2$. Slightly less close flybys with pericentre two or three times larger than r_T would be sufficient to dislodge the outer layers.

Exactly how frequently a star enters this 'zone of vulnerability' depends on the stellar velocity distribution. In a simple case when the velocities are isotropic at r_h, the frequency with which a solar type star would pass within a distance $r_{min}(\ll r_h)$ of the hole is

$$\approx 10^{-4}\, M_6^{4/3}\left(\frac{N_*}{10^5\ \mathrm{pc}^{-3}}\right)\left(\frac{\sigma}{100\ \mathrm{km\,s^{-1}}}\right)^{-1}\left(\frac{r_{min}}{r_T}\right)\ \mathrm{yr}^{-1} \qquad (2)$$

This 'fiducial' capture rate could be inaccurate for a number of reasons. The actual rate could be lower than that given by equation (2), even if the initial distribution were isotropic, if the near-radial 'loss cone' orbits were depleted faster than they could be replenished[18,19]. Replenishment could be slow if it occurred on the two-body star-star diffusion timescale, rather than by some collective instability or in a triaxial potential. If the relaxation timescale were short enough to ensure replenishment of the loss cone, then the same diffusion processes could build up a 'cusp' in the stellar distribution near the hole[17,18]; and a population bound to the hole could arise in other ways[20,21]. Capture or disruption of these stars could then augment the rate given by equation (2). If such a cusp were constituted by ordinary solar type stars, stellar collisions would overwhelm 'Coulomb' encounters within a radius GM_h/v_*^2, where $v_* = (GM_*/r_*)^{1/2}$. For holes of $<10^8\, M_\odot$, however, none of these complications is likely to seriously modify the simple estimate (2).

For hole masses $>10^7\, M_\odot$, complete disruption of solar-type stars occurs sufficiently close to the hole for a newtonian approximation to be inadequate. (Relativistic precession effects on orbiting debris are, moreover, important even for lower M_h, as mentioned later.)

The probability of a pericentric distance within r_{min} ($\ll r_h$) is proportional to the first power of r_{min} (see equation (2)). Complete disruption requires $r_{min} < r_T$. For r_{min} somewhat larger than r_T, a star could lose its outer layers, but the denser interior would survive: this intermediate possibility is specially important for giants. For a range of still larger values of r_{min} (but still, at least for main sequence stars, only exceeding r_T a few times), a solar-type star loses no material, but gets so distorted during the flyby that the subsequent internal dissipation (when it readjusts) might exceed its incoming kinetic energy $1/2\ m_*\sigma^2$; energy conservation then implies that the star would thereafter be on an orbit bound to the hole[22-24].

Specific binding energy of debris elements

The observational consequences of stellar capture and disruption depend on what happens to the debris. Consider first a solar-type star that passes somewhat closer to the hole than r_T and is disrupted in a single flyby. The energy required to tear the star apart (that is, the star's self-binding energy) is supplied at the expense of the orbital kinetic energy (which at r_T is larger by $\sim(M/m_*)^{2/3}$). Unless there were some explosive energy input, which would be expected only of the star passed several times closer than r_T (see ref. 25), on average the debris would be bound to the hole, unless the star was initially on a hyperbolic orbit with asymptotic velocity $> v_*$, which is $1{,}000\ \mathrm{km\,s^{-1}}$ for solar-type stars.

Several effects would, however, impart orbital binding energies spread widely about this mean to gas from different parts of the disrupting star; this spread crucially influences what we would actually observe. The dominant effect is the following. While falling inwards towards the hole, the star would develop a quadrupole distortion which attains an amplitude of order unity by the time of disruption at $r = r_T$. The resultant gravitational torque would 'spin it up' to a good fraction of its co-rotation angular velocity by the time it gets disrupted: it would consequently, by that stage, be spinning at close to its break-up angular velocity. The parts on the 'outside track' furthest from the hole would therefore have an extra velocity, over and above the orbital velocity $v_{orb} \simeq (2GM/r_T)^{1/2} \simeq c(r_g/r_T)^{1/2}$, of order $v_* = (m_*/M_h)^{1/3} v_{orb}$; those closest to the hole would have a comparable velocity deficit. Moreover, the slower-moving gas on the 'inside track' is deeper in the potential well by an amount $\sim(GM_h/r_T)(r_*/r_T) \simeq (Gm_*/r_*)$. There would consequently be a spread of order $v_{orb}\Delta v$, where $\Delta v \simeq v_*$, in the energies of different bits of debris. Other processes during the flyby—for instance impulses from shocks during the drastic compression and distortion of the stellar material—could further enhance Δv.

Even though the mean specific binding energy of the debris to the hole would be positive, and comparable with the self-binding energy (Gm_*/r_*) of the original star, the spread about this mean is larger by $(M_h/m_*)^{1/3}$—a factor which is >100 for hole masses M_h in the range relevant to galactic nuclei. Whenever a solar-type star passed within the tidal disruption radius r_T, some of the debris would be flung out on hyperbolic orbits with escape velocities up to $\sim10^4\ \mathrm{km\,s^{-1}}$. (This important qualitative point was apparently first appreciated by Lacy *et al.*[26], and discussed by Rees[27] in the context of our Galactic Centre). When the central hole has a mass $M_6 > 1$, the bound debris would be on orbits with characteristic specific binding energy $\gtrsim 10^{-3}\ c^2$ rather than $\sim10^{-5}\ c^2$. The actual mass-fractions in a particular energy range, and the precise high-and low-energy cut-offs, would depend on the details; when a specific assumption is needed for illustrative purposes I shall assume that the mass-fractions are uniformly distributed in energy throughout the range from E_{min} (bound to hole) to E_{max} (unbound), where $E_{min} = -(Gm_*/r_*)\,((M_h/m_*)^{1/3}+1)$ and $E_{max} = (Gm_*/r_*)\,((M_h/m_*)^{1/3}-1)$.

Timescale for swallowing bound debris

The average specific binding energy of debris bound to the hole (just over half the total mass, if we adopt the illustrative model for the energy distribution mentioned above, and assume $M_h \gg m_*$) is $\sim\frac{1}{2}(Gm_*/r_*)\,(M_h/m_*)^{1/3}$. The bound orbits are very eccentric; for solar-type stars the orbital major axis of even the most tightly-bound debris is $\sim10^3\, M_6^{-1/3}\, r_g$, and the period is only $t_i \simeq 0.03\, M_6^{1/2}$ yr: note that this is shorter by a factor $\sim10^3\, M_6^{1/2}$ than the orbital period if the specific binding energy were (Gm_*/r_*), as had been assumed in pioneering discussions of this topic (such as refs, 2, 28). For $M_6 < 100$, this is certainly small compared with the interval between successive disruptions (equation (2)). Unless it takes many orbital periods to swallow the bound gas, the debris from each star would be digested separately—in contrast to the debris cloud model of quasars[2], where the disruptions are more frequent but the orbital periods of the debris ($\propto M_h^{1/2}$) are longer. I shall now give a reason why the gas could not persist for more than very few orbits without being swallowed or expelled.

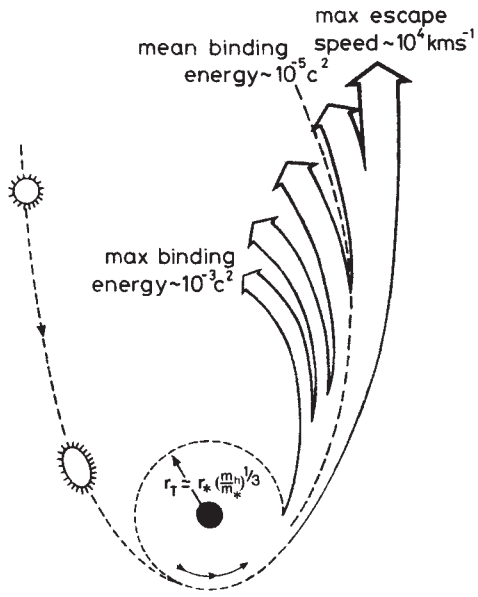


Fig. 1 A solar-type star approaching a massive black hole on a parabolic orbit with pericentre distance r_T is distorted and spun up during infall, and then tidally disrupted. The average specific binding energy of the debris to the hole is $\sim 10^{-5}\,c^2$ (of order the self-binding energy of the original star). However, the *spread* in this energy [of order $v\Delta v$, where $v \simeq c(r_T/r_g)^{-1/2}$ and $\Delta v \simeq (Gm_*/r_*)^{1/2}$] is $\sim 10^{-3}\,c^2$ for hole masses $M_h \gtrsim 10^6\,M_\odot$. Almost half the debris would therefore escape on hyperbolic orbits with speeds up to $\sim 10^4$ km s^{-1}; the most tightly bound debris would traverse an elliptical orbit with major axis $\sim 10^3\,r_g$ before returning to $r \simeq r_T$.

The range of orbital periods is large: if there were an equal mass-fraction in each interval of binding energy, then half the bound debris would have periods $>2^{3/2}$ longer than the innermost debris and a (small) mass fraction f would have periods longer by $>f^{-3/2}$. If the gaseous debris suffered no internal dissipation due to high viscosity or shocks, it would, after one or two orbital periods, form a highly elliptical disc with a big spread in apocentric distances between the most and least bound orbits, but where at pericentre the orbits were all squeezed in a range $(\Delta r/r) \simeq 0.01\,r_T$ for $M_6 \simeq 1$ (compare ref. 16). But we can readily see that a low-viscosity thin disc hypothesis is not tenable.

At each pericentre passage, where $r = r_T$, the relativistic orbital precession amounts to $\sim 3(r_g/r_T)$ radians, which for solar-type stars is $\sim 0.2\,M_6^{2/3}$ radians. This means that internal shocks with velocities of order $c(r_g/r_T)^{1/2}$ are inevitable once the most tightly bound gas has traced out more than one orbit. The debris raining down from 10^3–$10^4\,r_g$, all with specific angular momentum $\sim c r_T^{1/2} r_g^{1/2}$ appropriate to circular orbits at $r = r_T$, would, after little more than its free-fall time, settle into an axisymmetric torus, supported by radiation pressure, whose density maximum occurred in a ring with radius $\sim r_T$. This torus starts to form when the most tightly bound debris falls back in—at a time t_i after the disruptive flyby—and is thereafter fed by the injection of infalling matter at a rate that drops off as $t^{-5/2}$ for $t \gg t_i$ (again adopting the assumption of equal mass fractions in equal energy intervals). When a solar-type star has been disrupted by a close passage with pericentre distance $\sim r_T$, this infall rate is roughly given by

$$\dot{M} \simeq 25\,M_6^{-1/2}(t/t_i)^{-5/2} \quad \text{solar masses per year} \qquad (3)$$

The orbital period at r_T is a few hours ($\sim 10^{-3}$ yr). The viscous dissipation time for a thick torus would be α^{-1} orbital periods, α being the usual viscosity parameter; α would have to be below 10^{-3} for the dissipation time to exceed 1 yr. Such a low viscosity may be precluded by dynamical instability; moreover, the hole's spin axis would generally be misaligned with the star's orbital angular momentum, the resultant Lense-Thirring precession of the torus then being an extra cause of viscosity.

We then come up against another limit. The binding energy associated with a mass m_* in circular orbit at $r = r_T$ cannot be radiated away on a shorter timescale than $2.5 \times 10^8\,(m_*/M_h)(r_T/r_g)^{-1}$ yr without exceeding the hole's Eddington luminosity. This sets a minimum timescale of $\sim 5\,M_6^{-1/3}$ years for a solar mass of material to deflate into a thin ring at $r = r_T$, even if there were no further viscous dissipation, and $50\,\varepsilon_{0.1} M_6^{-1}$ yr for it to be swallowed by the hole with radiative efficiency $\varepsilon = 0.1\,\varepsilon_{0.1}$. Equivalently, the hole cannot accept matter, with radiative efficiency ε, at a rate exceeding

$$0.02\,\varepsilon_{0.1}^{-1} M_6 \quad \text{solar masses per year} \qquad (4)$$

The radiative efficiency depends on the viscous timescales in the inner axisymmetric flow pattern at $r < r_T$.

From equations (3) and (4) we readily infer that, if (as seems likely) the viscosity were high enough to process all the material within the infall timescale, then the luminosity L could not remain as high as the Eddington luminosity for longer than

$$17\,\varepsilon_{0.1}^{-0.4} M_6^{-0.6}\,t_i \sim 0.5\,M_6^{-0.1}\,\varepsilon_{0.1}^{0.4} \quad \text{yr} \qquad (5)$$

L would thereafter fall off as $t^{-5/2}$—indeed, the fall-off would probably be even steeper eventually , because the radiative efficiency of thermal processes in accretion flows generally falls off for very low $\dot{M}$; when $t \gg t_i$ the prime radiative output may be low-level non-thermal emission.

When the mass supply (3) exceeds the swallowing rate that (4) permits for high efficiency, there are two possibilities: (i) the material may all be swallowed, ε adjusting downwards so that all can be accepted without $\varepsilon \dot{M} c^2$ exceeding the Eddington luminosity. This could come about, in a manner familiar from studies of steady accretion tori, if a flow pattern established itself around the hole whose innermost parts followed orbits with binding energy close to zero. Material could then spill into the hole from such orbits with corresponding low radiative efficiency. (ii) Only a small fraction of the injected material may be swallowed, yielding radiation with high efficiency, the surplus being expelled in a radiation-driven wind.

One cannot (except by detailed considerations of the spectrum of the emergent radiation and so on) distinguish between these two possibilities. They do, however, differ importantly with regard to the rate at which the hole gains mass for a given stellar disruption rate (2), and the likelihood of a high-speed radiation-driven outflow.

A flare could persist steadily with $L \simeq L_E$ for longer than (5) only if the viscosity were so implausibly low that the bulk of the mass could be stored for many years in a 'reservoir' at $r = r_T$ (where the dynamical timescale is a few hours). In all other circumstances, the bulk of the debris would be swallowed or expelled rapidly compared with the interval between successive stellar captures—the only conspicuous luminosity being a flare (predominantly of thermal ultraviolet or X-ray emission) with peak intensity $L = L_E$, fading within a few years. The integrated output from this flare could, in principle, amount to a few per cent of the star's rest mass, but would probably be far less, because most debris would be 'fed' to the hole at $t = t_i$ far more rapidly than it could be accepted if ε were high; much would then escape in a radiatively-driven directed outflow or be swallowed inefficiently.

Unbound debris

When a star passes close enough to be disrupted in a single flyby, almost half the debris escapes on hyperbolic orbits. On the assumption (used above in connection with the bound orbits) that there is an equal mass-fraction in each energy interval up to E_{max}, the kinetic energy of the ejecta is $\sim 10^{51} M_6^{2/3}$ ergs. The material would be concentrated in a cone or 'fan' close to the

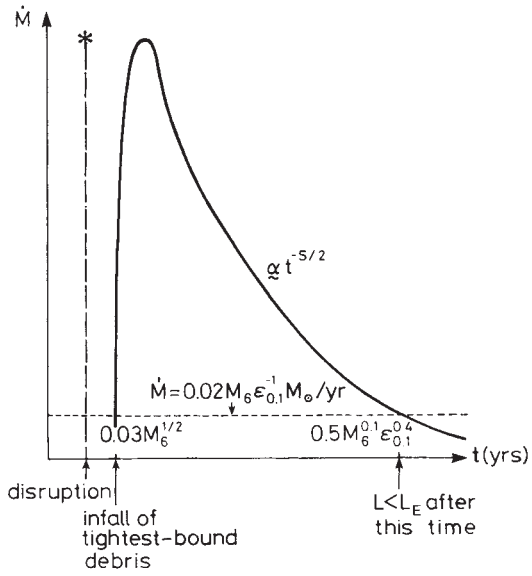


Fig. 2 When a disruptive flyby occurs with pericentric distance r_T, the bound debris starts to rain down on to the hole after a timelag $\sim 0.03\ (M_h/10^6\ M_\odot)^{1/2}$ years. This diagram shows schematically the subsequent behaviour of $\dot{M}$. The peak infall rate is highly 'super-Eddington'. This debris forms a torus at $r \simeq r_T$ within which radiation pressure is dominant. Its subsequent flow is controlled by viscosity, which is likely to operate on a timescale only a modest multiple of the rotation period at $\sim r_T$ (only a few hours). A thermal luminosity $\sim L_E$ can only be maintained for $\lesssim 1$ year; thereafter the 'flare' would rapidly fade.

orbital plane. Adiabatic cooling would severely reduce the internal radiative content before the debris became translucent. The 'prompt' radiation from the unbound debris would therefore release less than the initial energy content of the star (just as a supernova would be optically inconspicuous were it not for continuing energy injection in the months after the explosion). There would therefore be no conspicuous flare until, as discussed above, the bound debris fell back onto the hole after a time delay t_i. The main observable effects of the outflowing material would occur when it was decelerated by running into external diffuse matter.

Ultraclose passages and relativistic effects

The behaviour of stars passing well within the tidal radius r_T exhibits special features, some of which have been discussed in a series of paper by Carter, Luminet and collaborators[25,29-31]. Such stars are not only elongated along the orbital direction, but are even more severely compressed into a prolate shape (that is, a pancake aligned in the orbital plane). This compresson is halted by a shock, raising the matter (which then rebounds perpendicular to the orbital plane) to a higher adiabat. Also, there is the possibility of explosive energy release: the p-p reaction is too slow to release much energy on a dynamical timescale; however, proton capture on C, N and O can yield (for solar abundances) enough energy to unbind the star. The mean specific binding energy of the debris to the hole may then become negative, rather than being $\sim(Gm_*/r_*)$. But we have seen that, even in the latter case, about half the debris is unbound because of the spread $v_{orb}\Delta v$ in the energies at pericentre (where $\Delta v \simeq v_*$). If a nuclear explosion were to enhance Δv by some further factor, then the mass fraction escaping on hyperbolic orbits would still be ~ 0.5, but the range of energies of the debris would be increased: some would escape with higher hyperbolic velocities, whereas some would be even more tightly bound than before. To quantify this, one would need to know the angular distribution (in a frame sharing the star's mean motion) of the post-shock velocities, because velocity boosts perpendicular to the orbital plane yield only a second-order contribution, and are therefore less important than those in the plane.

The orbits for which extreme compression occurs would enter regions where relativistic effects were more important than for those with $r_{min} = r_T$. The orbits are then not ellipses, but may turn through 2π or even more. A test particle not exactly in the orbital plane (which in the case of an elliptical orbit would cross the plane just once at pericentre), may then have two or more traversals. There is then the possibility of multiple shocks[30,31].

For hole masses $M_h > 10^8\ M_\odot$, solar-type stars cannot be disrupted without entering the strongly relativistic domain. The form of the black hole (Schwarzschild or Kerr?) then has an important quantitative effect, as does (for a rotating Kerr hole) the orientation of the stellar orbit relative to the hole spin axis. Stars on counter-rotating orbits are more readily captured with the result that Kerr holes would spin down if they gained mass primarily from stellar capture[32]. When the hole mass is $\gg 10^8\ M_\odot$, most main sequence stars would be swallowed whole ($r_T < r_g$), and only giants would generate debris outside the hole[1,2].

Tidal dissipation without complete disruption?

A star which does not pass close enough to the hole to suffer complete disruption may nevertheless be tidally distorted; if the associated dissipation exceeds $m_*\sigma^2$, the star may then be captured in an elliptical orbit around the hole, and suffer further dissipation on each successive pericentric passage. This process has been primarily discussed in the context of globular clusters[22-24], where the star is of comparable mass to the compact object (stellar mass black hole or neutron star?); after capture, the orbit degrades and circularizes at a radius a few times r_T. One's first thought might therefore be that, in the present context, solar-type stars could be captured into near-circular orbits very tightly bound to a central massive hole. Such orbits at $r = r_T$ would have periods of a few hours (independent of M_h), and the orbital velocity would be $\sim 0.2\ M_6^{1/3}\ c$—vastly higher, of course, than the speeds involved in ordinary binary systems.

But there is a crucial new aspect to this tidal-capture mechanism when a very massive black hole is involved. In acquiring a tightly bound circular orbit of radius r_T, the star must dispose of an amount of energy $m_*c^2(r_T/r_g)^{-1}$—larger by a factor $(M_h/m_*)^{2/3}$ than its entire original gravitational self-binding energy Gm_*^2/r_*. The circularization would therefore need to occur much slower than the star's thermal timescale. Of course, the star would, after the first few passages, acquire a spin rate synchronized with its orbital angular velocity at pericentre, and thereafter the dissipation could be reduced somewhat below the rates calculated in refs 22-24.

Until appropriately detailed three-dimensional gas-dynamical calculations can be done, the fate of such stars will be an open question. It would seem more likely that, after the orbit had been 'ground down' by the first few passages (and the frequency of subsequent pericentre passages had become correspondingly higher), the star would inflate and get fully disrupted. The gaseous debris would then evolve essentially as the bound fraction of the debris from the tidally disrupted stars discussed earlier. More uncertain would be the fate of a giant star, where the dissipation, occurring primarily in the envelope, could leave the core relatively undistorted and unscathed. (If internal dissipation in the resultant torus were exceedingly low, the debris could deflate on the timescale $\sim 5\ M_6^{-1/3}$ years into a thin ring at $r = r_T$ which could in principle be vulnerable to gravitational instability, leading to rebirth of a star (compare ref. 33). But this would require an effective viscosity parameter $< 10^{-4}$, which is perhaps implausibly low.)

There is one further interesting mechanism for injecting stars into small tightly-bound orbits (though not necessarily as small as $r = r_T$. Diffusive stellar-dynamical processes could not inject stars into orbits with specific binding energies exceeding v_*^2—at such velocities disruptive stellar collisions dominate 'Coulomb' encounters. But a fraction of the stars in the core of the galaxy

might well be binaries: tidal disruption of a binary by the hole, equivalent to a three-body exchange process, could (for reasons analogous to those already discussed in the context of the stellar debris) eject one member of the original binary on a high-speed hyperbolic orbit, leaving the other in an orbit tightly bound to the hole. Three-body calculations of this process have recently been presented by Hills[21]. The pericentre for such stars would be of order the 'tidal radius' for the original binary, and therefore well outside the tidal-disruption radius r_T for an individual star. Subsequent tidal circularization would therefore not occur fast enough to trigger disruption, and some slow frictional process (such as drag on gas) could then perhaps grind down the orbit until it approached $r = r_T$.

Despite the uncertainties (and perhaps improbabilities), the processes mentioned above indicate that it is far from inconceivable for a solar-type star to become established in an orbit near $r = r_T$. Its gravitational radiation decay time would then be

$$t_{GR} \simeq 10^6 M_6^{-2/3} \text{ yr} \quad (6)$$

typically long compared with the interval (equation (2)) between successive captures. Drag due to interaction with gas may degrade the orbit rather faster, but any star that did manage to get into a tightly bound orbit near $r = r_T$ could remain there until another star got captured on to (or settled into) a similar orbit. Any such star might be hard to detect directly, even in our own Galactic Centre (though an ordinary star with an anomalously high orbital velocity would offer gratifyingly unambiguous evidence for a massive black hole there). In an active galactic nucleus, such a star might modulate the intensity of any radiation due to gas accretion or electromagnetic effects. The search for strictly periodic variability in continuum emission from galactic nuclei, though a long shot, offers sufficient payoff to merit some continuing effort. Such stellar orbits woud display interesting relativistic effects. For instance, if the central hole were spinning, and the star's orbital angular momentum were misaligned with respect to the hole's spin axis, then Lense-Thirring precession of the orbital plane would have a timescale of $(J/J_{max})^{-1}$ (r_g/c) $(r_T/r_g)^3$—only of the order of months for parameters relevant to the Galactic Centre, if the hole's angular momentum J were comparable with J_{max}, the maximum attainable by a Kerr metric.

A black hole at the Galactic Centre?

The dynamical evidence for a massive black hole at our Galactic Centre (reviewed in ref. 13) is based primarily on interpretation of gas motions within the central parsec. These data, together with the unresolved radio source[12], increasingly suggest that there is some unique object right at the centre, even if much of the activity and structure outside the central parsec can be accounted for in conventional terms. The situation is, however, confused by the fact that the brightest part of the central infrared source IRS 16 appears displaced from the radio source[34].

Because of intervening obscuration, we do not directly know the velocity dispersion of stars within the central parsec of our Galaxy; observations of the gas do, however, tell us something about the total mass in that region. Bailey[35] showed that, if the entire central mass concentration were a star cluster, then at a radius 0.1 pc the stars would be so closely packed that runaway dynamical evolution would take $<10^9$ yr. Unless we are observing at a privileged epoch in the history of our galactic nucleus, this supports the idea that a massive central object may already have formed.

The black hole hypothesis receives at least circumstantial support from the new evidence for black holes in nearby galaxies[5-9]. Perhaps all spiral galaxies pass intermittently through Seyfert phases: the black hole could then exist as a relic of earlier Seyfert-style outbursts, being reactivated when the fuelling rate was boosted. Recent evidence that the compact central radio source has a transverse motion <40 km s^{-1} (ref. 36) suggest that this source, whatever it is, indeed lurks right at our Galaxy's dynamical centre. If a massive hole were present in the central regions of any galaxy, then dynamical friction would settle it quickly into the true dynamical centre.

The small central radio component has an unusual (rising) spectrum; if there were a central hole with $M_6 = 1$, then the radio emission could come from plasma participating in a low-level accretion flow and gravitationally bound to the hole[33]: tenuous plasma would be unable to cool on the inflow timescale, the only substantial radiative output coming from within 1,000 Schwarzschild radii, where even the average electron could be relativistic with energy $\gtrsim 1$ Mev. There are, admittedly, alternative models that do not require a massive hole[37-39]. On the other hand, a massive hole accreting at a low rate would almost inevitably yield radio emission: the absence of a peculiar radio source at the Galactic Centre would therefore have been evidence against such an object there.

It would be a coincidence (at more than the 1% level) if we were now observing the Galactic Centre just at the time of a flare or stellar disruption. There is therefore no problem with the hypothesis that a star is disrupted once every $\sim 10^4$ yr, part being swallowed and part ejected—it is certainly acceptable that the mean luminosity from the centre should exceed what we now observe. Special interest attaches to ths particular galactic nucleus not only because it is 'close to home' but because it may exhibit residual traces of individual stellar-disruption events. One speculative possibility is the following.

The volume within ~ 2 pc of the centre is probably pervaded by gas at 3×10^6 K with density ~ 100 cm^{-3}. This gas, with cooling time $\sim 3 \times 10^4$ yr could, if there were indeed a central hole of $\sim 10^6 M_\odot$, be heated by the ejecta from disrupted stars[40]. The stellar debris (enriched in helium) would contribute to the energy budget within the central 2 pc, if not to the mass supply. The mass of fast-moving ejecta would be small, but, if its density were high, it might be relatively conspicuous. It would be interesting to consider its possible relevance to the observed fast-moving He feature.

Moreover, the peculiar arm-like features revealed by the Ne line and a thermal radio continuum[11,26] consist of denser gas with temperature 10^3–10^4 K. The cooling of arm-like swaths of gas could have been triggered by the rapid directed outflow from the most recent stellar disruption events (which might have occurred $<10^4$ yr ago). If this interpretation (discussed further in refs 27 and 40) were correct, these features could be 'vapour trails' delineating the tracks of fast ejecta from individual disrupted stars. The cool gas would not itself be ejected material; it would be ambient (perhaps slowly infalling) gas in the cavity whose cooling was triggered by the overpressure. The situation would resemble a 'sector' of a supernova remnant.

Relevance to the nuclei of nearby galaxies

The most obvious consequence of a 10^6–10^8 $M_\odot$ black hole would be transient flares whenever bound debris from a star was swallowed. The rate is given by equation (2) with $r_{min} = r_T$, the luminosities being as high as $L_E = 10^{44} M_6$ erg s^{-1}. But in a given object, these flares would have a duty cycle of $\sim 10^{-3}$ at peak luminosity (equation(5)). After each luminosity peak, there would be a decline as the 'dregs' were swallowed, but the median luminosity would be far below that which would result from steady efficient accretion of the mass supply implied by (2). Therefore we would not yet expect to have detected such a flare. On the other hand, a sufficiently large sample of such galaxies should reveal some members of the ensemble in a flaring state. Such objects could be searched for out to large distances: they would differ from typical AGNs through the lack of any extended structure (emission line or radio components). If $\sim 10^6 M_\odot$ holes were prevalent in small or even dwarf galaxies, the nearest such flares, in any given year, may be no further away than the Virgo Cluster.

The mass fraction that is ejected rather than swallowed, though less spectacular than the accretion-powered flares, could

nevertheless have cumulative effects. There are two ways in which matter can be expelled: (a) When a star is disrupted in a single flyby, about half the debris is ejected in a fast-moving spray of gas. The characteristic velocities are $\sim 10^4\ M_6^{1/6}$ km s^{-1}, the kinetic energy output when a solar-type star is disrupted being $\sim 10^{51}\ M_6^{1/3}$ erg s^{-1}—for $M_6 > 1$ (that is, in all cases of interest here), this exceeds the energy of a supernova. Stars that penetrate well inside r_T would cause still more violent ejection. (b) The bound debris falls back when $t > t_i$ to $r \approx r_T$ where it can acquire centrifugal support. If energy is dissipated at a supercritical rate as ths material swirls closer to the hole, some gas may then be ejected in a radiation-driven wind, or even in double jets aligned with the angular momentum of the debris. In principle it is possible for only a fraction $\varepsilon^{-1}(r_g/r_T)$ to be actually swallowed, ε being the efficiency, all the remainder being ejected. The characteristic speeds here—of order the escape velocity from r_T—would be even higher than those of type (a) outflow discussed above. The debris from a star that does not approach close enough to be destroyed in a single flyby but is eventually disrupted by tidal dissipation could behave similarly.

If there were indeed a $\sim 10^6\ M_\odot$ hole in our Galactic Centre, the consequences of the unbound (outgoing) debris may be conspicuous—indeed, I conjectured above (and in refs 27 and 40) that they may have been unknowingly observed already. The direct effects of unbound debris from an individual disruption would not be readily discernable, however, even in a nearby galaxy such as M31 or M32. The fast-moving ejecta from successive disruptions could nevertheless have an important cumulative influence on the energy balance within the central 10-100 parsecs of a galaxy—more, quite probably, than supernovae exploding in that same volume. The ejecta would be braked, and their kinetic energy thermalized, as they ran into the diffuse gas originating from stellar mass loss or infall from the body of the galaxy. X-ray observations offer direct evidence that there is hot gas in elliptical galaxies, with pressure $nT = 10^7$ cm^{-3} K in the central kiloparsec[41]. Steady accretion of this gas even at the minimum 'Bondi' rate (neglecting cooling) would yield conspicuous effects unless the radiative efficiency were very low[42,43].

The accretion radius is $1.5 \times 10^{17}\ M_6\ T_7^{-1}$ cm. If accretion proceeded unimpeded, the energy within this radius (taking the foregoing estimate of the ambient pressure) would be $10^{43}\ M_6^{-3}\ T_7^{-3}$ erg: even for M_6 as large as 100, this amounts to less than the energy from a single disrupted star. For $M_6 = 100$, the characteristic inflow timescale would be $1.5.10^4\ T_7^{-3/2}$ yr. Comparison with the stellar capture rate (2) confirms that this timescale is long enough for the ejected debris to be equivalent to a fast wind: its ram pressure and heat input could inhibit the steady accretion that would otherwise be inevitable in any galaxy harbouring a black hole.

Stellar disruption and coalescence may have other implications—for instance, the fast moving ejecta may be relevant to the dynamics of line-emitting regions. But they could definitely have the paradoxical effect of making the average nuclear luminosity even lower than would be estimated if only gas were present, because the fast-moving ejecta could stem the steady accretion flow. Thus the quiescence of a nearby nucleus is not incompatible with the presence within it of a massive black hole of $10^8\ M_\odot$.

This paper has outlined several potentially observable effects that call for detailed stellar-dynamical and hydrodynamic calculations. Precise modelling should also allow for a realistic spread in the stellar population. The stellar disruption phenomenon is relevant to high-resolution observations of nearby galaxies, and to the properties of AGNs in general. In the meantime, it would seem worthwhile to search for thermal flares lasting ≤ 1 yr in the nuclei of ordinary quiescent galaxies.

I am grateful for helpful comments from A. C. Fabian, H-M. Lee, J. Luminet, R. Nityananda, W. Saslaw and C. H. Townes.

Received 5 January; accepted 21 March 1988.

1. Hills, J. G. *Nature* **254**, 295-298 (1975).
2. Hills, J. G. *Mon. Not. R. astr. Soc.* **182**, 517-530 (1978).
3. Frank, J. *Mon. Not. R. astr. Soc.* **184**, 87-99 (1978).
4. Young, P. J., Shields, G. A. & Wheeler, J. C. *Astrophys. J.* **212**, 367-382 (1977).
5. Tonry, J. *Astrophys. J.* **283**, L27-L30 (1984).
6. Dressler, A. *Astrophys. J.* **286**, 97-105 (1984).
7. Kormendy, J. in *Elliptical Galaxies* (ed. de Zeeuw, T.) 17-36 (Reidel, Dordrecht, 1987).
8. Kormendy, J. *Astrophys. J.* **325**, 128-141 (1988).
9. Dressler, A. & Richstone, D. O. *Astrophys. J.* **324**, 701-713 (1988).
10. Lynden-Bell, D. & Rees, M. J. *Mon. Not. R. astr. Soc.* **152**, 461-475 (1971).
11. Serabyn, E. & Lacy, J. H. *Astrophys. J.* **293**, 445-458 (1985).
12. Lo, K. Y. *Science* **233**, 1394-1400 (1986).
13. Genzel, R. & Townes, C. H. *A. Rev. Astr. Astrophys.* **25**, 377-423 (1987).
14. Ozernoi, L. M. in *Large Scale Characteristics of the Galaxy* (ed. Burton, W. B.) 395-400 (Reidel, Dordrecht, 1979).
15. Gurzadyan, V. G. & Ozernoi, L. M. *Astr. Astrophys.* **86**, 315-320 (1980).
16. Gurzadyan, V. G. & Ozernoi, L. M. *Astr. Astrophys.* **95**, 39-45 (1981).
17. Bahcall, J. N. & Wolf, R. A. *Astrophys. J.* **209**, 214-232 (1976).
18. Frank, J. & Rees, M. J. *Mon. Not. R. astr. Soc.* **176**, 633-647 (1976).
19. Lightman, A. P. & Shapiro, S. L. *Astrophys. J.* **211**, 244-262 (1977).
20. Young, P. J. *Astrophys. J.* **242**, 1232-1237 (1980).
21. Hills, J. G. *Nature* **331**, 687-689 (1988).
22. Fabian, A. C., Pringle, J. E. & Rees, M. J. *Mon. Not. R. astr. Soc.* **172**, 15P-18P (1975).
23. Press, W. H. & Teukolsky, S. A. *Astrophys. J.* **213**, 183-192 (1977).
24. Lee, H-M. & Ostriker, J. P. *Astrophys. J.* **310**, 176-188 (1986).
25. Carter, B. & Luminet, J. P. *Nature* **296**, 211-214 (1982).
26. Lacy, J. H., Townes, C. H. & Hollenbach, D. J. *Astrophys. J.* **262**, 120-134 (1982).
27. Rees, M. J. in *The Milky Way Galaxy* (eds van Woerden, H. *et al.*) 379-383 (Reidel, Dordrecht, 1985).
29. Frank, J. *Mon. Not. R. astr. Soc.* **187**, 883-904 (1979).
29. Luminet, J. P. & Carter, B. *Astrophys. J. Suppl.* **61**, 219-248 (1986).
30. Luminet, J. P. in *Gravitation in Astrophysics* (eds Carter, B. and Hartle, J. B.) 215-228 (Reidel, Dordrecht, 1987).
31. Luminet, J. P. & Marck, J.-A. *Mon. Not. R. astr. Soc.* **212**, 56-75 (1985).
32. Young, P. J. *Astrophys. J.* **212**, 227-233 (1977).
33. Rees, M. J. in *The Galactic Center* (eds Riegler, G. R. & Blandford, R. D.) 166-176 (American Institute of Physics, New York, 1982).
34. Allen, D. A. & Sanders, R. H. *Nature* **319**, 191-194 (1986).
35. Bailey, M. E. *Mon. Not. R. astr. Soc.* **190**, 217-225 (1980).
36. Backer, D. C. & Sramek, R. A. in *The Galactic Center* (ed. Backer, D. C.) 163-165 (American Institute of Physics, New York, 1987).
37. Reynolds, S. P. & McKee, C. F. *Astrophys. J.* **239**, 893-897 (1980).
38. Kardashev, N. S., Novikov, I. D., Polnarev, A. G. & Stern, B. E. in *Positron-Electron Pairs in Astrophysics* (eds Burns, M. L. *et al.*) 253-266 (American Institute of Physics, New York, 1985).
39. Ozernoi, L. M. in *The Galactic Center* (ed. Backer, D. C.) 181-183 (American Institute of Physics, New York, 1987).
40. Rees, M. J. in *The Galactic Center* (ed. Backer, D. C.) 71-78 (American Institute of Physics, New York, 1987).
41. Thomas, P. *Mon. Not. R. astr. Soc.* **220**, 949-969 (1986).
42. Begelman, M. C. *Nature* **322**, 614-615 (1986).
43. Fabian, A. C. & Canizares, C. R. *Nature* (submitted).

Diabetes in transgenic mice resulting from over-expression of class I histocompatibility molecules in pancreatic β cells

Janette Allison, I. L. Campbell, G. Morahan, T. E. Mandel, L. C. Harrison & J. F. A. P. Miller

The Walter and Eliza Hall Institute of Medical Research, PO The Royal Melbourne Hospital, Victoria 3050, Australia

A class I histocompatibility gene, H-2K^b, linked to the rat insulin promoter, is overexpressed in the pancreatic β cells of transgenic mice. The mice, whether syngeneic or allogeneic to the transgene, develop insulin dependent diabetes without detectable T cell infiltration, suggesting a direct, non-immune role for the transgenic class I molecules in the disease process.

BOTH human and animal models of insulin dependent diabetes mellitus (IDDM) suggest immunological involvement in its pathogenesis. In humans, HLA antigens encoded by the major histocompatibility complex (MHC) determine disease susceptibility[1], but the precise mode of action of the MHC has not been defined. Although multiple, varied and sometimes transient immunological abnormalities have been reported in patients with diabetes, it is still not clear whether immune factors are causative or contributory, or whether they result from destruction of the insulin-producing β cells of the pancreas[2].

A transgenic mouse model provides a way of testing the hypothesis that immune factors are responsible for the development of diabetes. It is currently believed that tolerance of T cells to self antigens is limited to those molecules presented in association with MHC components within the thymus[3]. If so, one would expect diabetes to occur when a gene coding for a foreign MHC molecule was expressed exclusively in the pancreatic β cells. This model should allow us not only to test the autoimmune pathogenesis of IDDM, but also to investigate the role of the thymus in immune tolerance. Although the mechanism of T cell tolerance induction is not clear, there is persuasive evidence for the clonal elimination of T cells encountering self-MHC within the thymus[4]. There is, however, no evidence for T cell tolerance to circulating MHC molecules of the class I type[5].

Our strategy was to use the transgenic approach to direct the expression of a particular class I MHC molecule exclusively to the pancreatic islet β cells in strains of mice carrying the same or a different MHC haplotype. These cells normally express a low level of class I molecules[6]. We expected that a nonself class I molecule would provoke a T cell immune response leading to the development of IDDM. Diabetes did in fact develop in these mice at an early age. We were surprised, however, to find that the disease occurred in mice regardless of whether they were syngeneic or allogeneic with respect to the transgene product and in the absence of T cell involvement. This suggests a direct, non-immune, role for the transgenic class I molecules in the impairment of β cell function.

Diabetic transgenic mice

Hanahan[7] has shown that the upstream regulatory sequences of the rat insulin promoter can direct the expression of the simian virus 40 (SV40) large T antigen solely to the pancreatic islet β cells. Hence, to direct MHC class I expression to these cells, we have linked the H-2K^b protein coding sequences to the rat insulin promoter (Fig. 1) and microinjected this construct into fertilized eggs of mouse strains of H-2 haplotypes either syngeneic (b) or allogeneic (k, s) to the transgene. Initially,

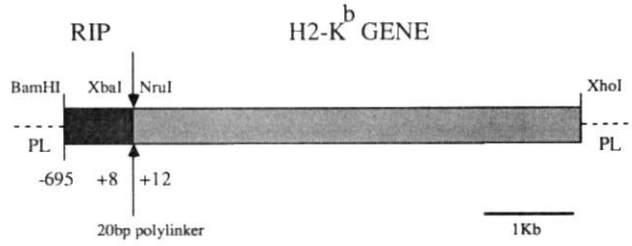


Fig. 1 The rat insulin promoter/H-2K^b construct. The H-2K^b gene of C57BL/10 origin was linked to the rat insulin promoter as follows. The *Bam*HI-*Xba*I fragment of pRIP1-Tag[7] containing the rat insulin II promoter (RIP) from −695 to +8 was subcloned into the *Bam*HI-*Xba*I site of pIC-20H[35] to give the vector pIC-RIP. The B10.5 kilobase *Eco*RI fragment containing the H-2K^b gene[36] was subcloned into pBSV. From this vector the H-2K^b *Nru*I-*Eco*RI fragment was subcloned into the *Sma*I-*Eco*RI site of pUC18. A *Sal*I-*Xho*I fragment containing the H-2K^b gene from +12 to the *Xho*I site downstream of the gene[37,38] was isolated and ligated to a *Sal*I digest of the pIC-RIP vector. The recombinant containing the appropriate insert orientation was selected and the bacterial DNA sequences were removed using the *Hin*dIII sites located in the polylinker (PL) at either end of the RIP-H-2K^b construct. This DNA was microinjected into fertilized mouse eggs[39]. The mouse strains used were B10.BR, (B10.BR × B10)F_1 and (C57BL/6 × SJL)F_2.

transgenic mice were made using the B10 congenic mouse strain in order to permit immunological studies using strains with the B10 background. Three transgenic B10.BR mice (haplotype kk and hence allogeneic to the transgene product) and one (B10.BR × B10)F_1 mouse (haplotype kb and hence genetically tolerant of the transgene product) developed IDDM, as defined by persistent glucosuria from an early age (Table 1). The urine glucose concentrations in these mice were greater than 110 mM. The mice were markedly underweight; it was difficult to control their hyperglycaemia with insulin injections and breeding attempts were unsuccessful. The occurrence of diabetes in strains of mice syngeneic with respect to the H-2K^b transgene product was unexpected and suggested that the hyper-expression of class I MHC molecules in pancreatic islet β cells might alone be responsible for the disease process.

Because of problems in mating and breeding of the B10 congenics, we used (C57BL/6 × SJL)F_2 hybrid mice (haplotypes bb, bs or ss). The mice developing from the (C57BL/6 × SJL)F_2 eggs may thus be syngeneic (and hence genetically tolerant) or allogeneic with respect to the H-2K^b transgene product. Six

Table 1 Transgenic founders carrying the RIP-H-2K^b gene construct

Mouse	Strain and haplotype	Onset of diabetes (days)	Glucose (mM) Urine	Blood*
35-1 ♀	B10.BR (kk)	<3	30	35
36-2 ♂	B10.BR (kk)	<20	>110	ND
40-3 ♂	B10.BR (kk)	<14	>110	ND
53-1 ♂	(B10.BR × B10)F_1 (kb)	<8	>110	ND
31-2 ♀	(C57BL/6 × SJL)F_2 (bs)	—	normal	
32-1 ♀	(C57BL/6 × SJL)F_2 (bb)	—	normal	
37-5 ♀	(C57BL/6 × SJL)F_2 (bb)	—	normal	
37-8 ♂	(C57BL/6 × SJL)F_2 (bs)	—	normal	
38-4 ♂	(C57BL/6 × SJL)F_2 (bs)	—	normal	
39-12 ♂	(C57BL/6 × SJL)F_2 (bb)	—	normal	

Transgenic founders from mouse strains B10.BR, (B10.BR × B10)F_1 and (C57BL/6 × SJL)F_2 were analysed for the RIP-H-2K^b transgene by Southern blotting of tail DNA using a *Bam*HI restriction digest and the RIP *Bam*HI-*Xba*I fragment as a DNA probe. Founders of the B10.BR strain developed diabetes at an early age and were too sick to breed. Founders of the (C57BL/6 × SJL)F_2 mice were still normal at 30 weeks of age. Haplotyping of founders developing from (C57BL/6 × SJL)F_2 eggs was determined by flow cytometry of peripheral blood lymphocytes stained with the anti-H-2K^b monoclonal antibody B8-24-3[33] or with a polyclonal C57BL/6 anti-SJL serum and visualized with an FITC-sheep anti-mouse Ig diluted 1:100 (Silenus). Tissue typing was confirmed by Southern blot analysis of restriction site polymorphisms between the b and s haplotypes using a *Bam*HI restriction digest and the 640 base pair (bp) *Bam*HI fragment upstream of the H-2K^b promoter as a DNA probe[34]. From these studies, we concluded that peripheral blood lymphocytes did not express the H-2K^b transgene. Onset of diabetes was detected by glucosuria using Diastix strips (Ames, Miles Laboratories, Australia). Blood glucose was determined by placing a drop on a BM-test-glycemie test strip (Boehringer Manheim). After 1 min, excess blood was wiped off and the glucose concentration measured with a Reflolux 11 reflectometer.

* Normal blood glucose is 10–18 mM for B10.BR and 5–10 for (C57BL/6 × SJL)F_2. ND, not determined.

transgenic founders developing from these eggs were of bb or bs haplotype (Table 1). None developed diabetes and are still normal at 30 weeks of age.

The founders of the (C57BL/6 × SJL)F_2 were mated to normal SJL mice (haplotype ss) and the offspring analysed for haplotype and presence of the transgene (Table 2). Three founders (31-2, 32-1 and 37-5) gave rise to offspring of bs or ss haplotypes; all transgenic offspring, regardless of haplotype, became diabetic between 18 and 25 days of age. Two other founders (38-4 and 39-12), however, did not give rise to any diabetic transgenic offspring.

The failure of all (C57BL/6 × SJL)F_2 transgenic founders to develop diabetes may be due to transgene mosaicism allowing enough normal β cells to develop[8] to maintain the mice in a non-diabetic state. The occurrence of diabetes in the B10.BR primary transgenic mice may be due to non-mosaicism or to an inordinate strain susceptibility to the effect of the transgene product. Transgenic offspring of the (C57BL/6 × SJL)F_2 founders cannot be mosaic and in three of the lineages (31-2, 32-1 and 37-5) they developed diabetes. The failure of transgenic offspring from two other lineages (38-4 and 39-12) to develop diabetes may result from an integration site position effect or transgene rearrangement that prevents gene expression. This pattern of transmission or non-transmission of diabetes for the five (C57BL/6 × SJL)F_2 lineages is absolute and has been analysed over three generations in over 100 mice.

Pancreas histopathology

Diabetic B10.BR mice were sacrificed at 30–40 days of age, by which time they were very debilitated. Histological sections of Bouin-fixed pancreas revealed some atrophy of the acinar cells but markedly abnormal islets (Fig. 2). These were fewer in number, depleted of β cells and had a disorganized structure in which the glucagon-producing α and somatostatin-producing δ cells, normally in the periphery, were intermingled with the remaining β cells. The α and δ cells, which would not be expected to express the transgene[7], did not appear decreased in number. There was no lymphocytic infiltration at this stage. Tissues other than the islets were normal except for the effects of diabetes-associated dehydration. They also had no lymphocyte infiltrate.

H-2K^b expression

Offspring of the (C57BL/6 × SJL)F_2 founders were sacrificed at different ages to determine the presence of H-2K^b and insulin in the islets by conventional immunohistochemical methods (Fig. 3). H-2K^b was detected as early as 2 days after birth in transgenic offspring of the 31-2 lineage. Although expression of H-2K^b was not detectable by the immunoperoxidase technique in non-transgenic bs mice, low level surface expression was revealed on isolated β cells analysed by immunofluorescence and flow cytometry. This level was increased on the β cells from the transgenic mice (Fig. 4). The islets of day 2 and day 14 transgenic mice appeared normal, but by day 30 insulin staining had decreased (Fig. 3). By day 60, very few islets remained (not shown) and mice usually died at this age unless given insulin injections. The islets of diabetic mice from the (C57BL/6 × SJL)F_2 diabetic line were not as abnormal as those of the B10.BR diabetics of comparable ages, and we were able to maintain males from the former strain on daily injections of isophane insulin for several months.

The immunoperoxidase technique failed to detect H-2K^b expression in the islets of transgenic offspring from a non-diabetic line (39-12 ♂), either at 2 days of age (not shown), or at 57 days (Fig. 3). The islets appeared completely normal and stained well for insulin. From this we conclude that H-2K^b gene

Table 2 Transgenic offspring of the (C57BL/6 × SJL)F_2 founders backcrossed to normal SJL mice

Founder and haplotype	Transgene copies transmitted	No. of offspring	Transgenic mice: Total	No. diabetic	Haplotype	Diabetes onset (days)	Glucose (mM) Urine	Blood
31-2 ♀ (bs)	4–6	19	5	5	3 bs, 2ss	18–25	>110	>44
32-1 ♀ (bb)	3–4	28	7	7	7 bs	18–25	>110	>44
37-5 ♀ (bb)	5–7	7	2	2	2 bs	18–25	>110	>44
38-4 ♂ (bs)	2–3	6	2	0	2 ss	—	normal	
39-12 ♂ (bb)	2–3	6	3	0	3 bs	—	normal	

Transgenic founders were mated to normal SJL mice (ss) and the offspring analysed for haplotype and presence of the transgene. The 31-2, 32-1 and 37-5 lineages have transmitted diabetes to all of 60 other transgenic offspring of the second and third generations so far analysed. The other two lineages have not yet given rise to any diabetic offspring.

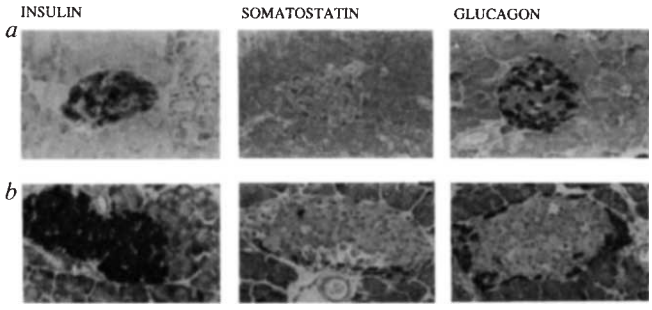


Fig. 2 Histopathology of pancreatic islets in a diabetic mouse. The B10.BR transgenic founder, 36-2, and a non-transgenic B10.BR mouse were killed at age 37 days and the pancreata were analysed for the presence of insulin, somatostatin and glucagon: *a*, the diabetic transgenic B10.BR mouse; *b*, the normal non-transgenic B10.BR mouse. The islets of the diabetic mouse stain poorly for insulin, are smaller in number and show a disorganized structure in comparison to the islets of the normal B10.BR pancreas. This disorganized structure is more apparent in the sections stained for somatostatin and glucagon. There are no detectable lymphocytic cells. Magnification ×150.

Methods. The pancreata were fixed in Bouin and embedded in paraffin. Immunohistochemistry was performed as described[40]. Insulin was detected using a two-stage immunoperoxidase procedure with guinea pig anti-porcine insulin antibody diluted 1:500 (ICN), visualized with a peroxidase conjugated rabbit anti-guinea pig Ig diluted 1:20 (Dako). Somatostatin and glucagon were detected using the peroxidase anti-peroxidase procedure. Rabbit anti-human somatostatin or glucagon diluted 1:200 (Dako) was followed by a swine anti-rabbit Ig diluted 1:20 (Dako) and visualized by a rabbit antibody against horse raddish peroxidase diluted 1:100 (Dako).

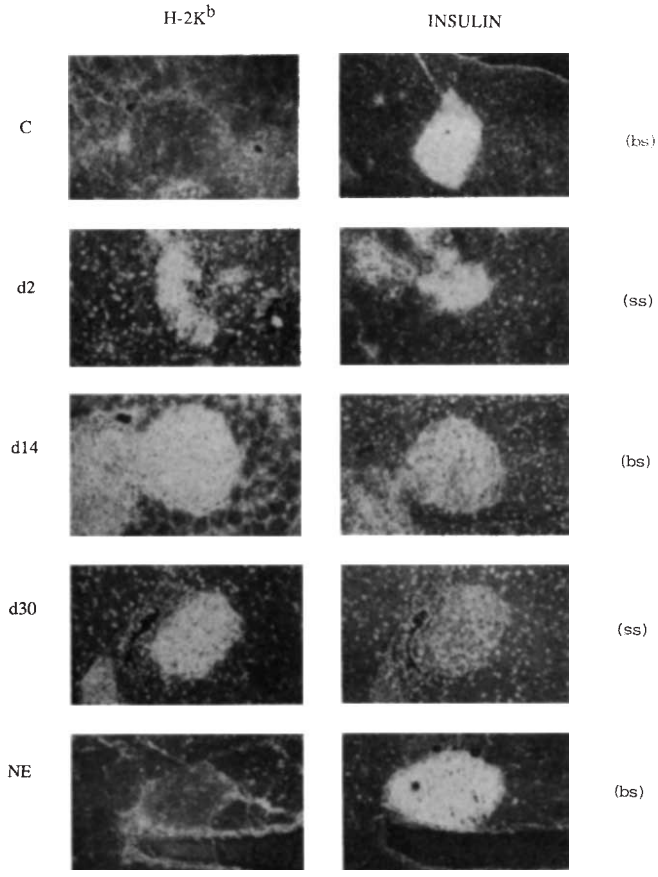


Fig. 3 Immunohistochemical detection of the H-2K^b protein in the islets of transgenic mice. The pancreata of diabetic (C57BL/6×SJL)F_2 transgenic offspring (lineage 31-2) and non-transgenic littermates were analysed for H-2K^b and insulin expression at days 2, 14 and 30 after birth using a two-stage immunoperoxidase assay. Top panel, control (C), day 14 non-transgenic pancreas. Second panel, day 2 pre-diabetic transgenic pancreas; third panel, day 14 pre-diabetic transgenic pancreas; fourth panel, day 30 diabetic transgenic pancreas. H-2K^b expression is present as early as day 2; insulin expression decreases as the disease progresses. The haplotypes of the individual mice, shown in brackets, were determined from the staining pattern of thymus or spleen tissue located in the same section. Islets of nontransgenic littermates aged 2 and 30 days (ss haplotype) stained normally for insulin (not shown here). Bottom panel, pancreas of a (C57BL/6×SJL)F_2 transgenic mouse (lineage 39-12) analysed at 57 days. H-2K^b expression is not detected in this mouse with the procedure used here. In contrast with the transgenic pancreata that do express H-2K^b, the islets appear normal and stain well for insulin. Dark field microscopy is used to enhance the contrast between stained and non stained areas. Magnification ×100.

Methods. Pancreata of offspring from the 31-2 lineage of the (C57BL/6×SJL)F_2 mice, or of the 39-12 lineage, were embedded in O.C.T. (Tissue Tek) and frozen in a dry ice/hexane bath. Sections (5 μM) were cut with a cryostat microtome, transferred to gelatin-coated slides and fixed in acetone for 10 min at 4 °C. They were stored at −70 °C until needed and then thawed for 1 min in acetone at room temperature. Staining for H-2K^b was performed by using the H-2K^b specific monoclonal antibody B8-24-3[33], visualized with a peroxidase conjugated sheep anti-mouse Ig diluted 1:200 (Silenus). Staining for insulin was as described in Fig. 2. Tissues were also stained with a negative contol monoclonal antibody specific for H-2K^k (11.4.1, ref. 41). As expected, no staining was observed with this antibody (not shown).

expression is a requirement for the disease process. The presence of a few tandem copies of the rat insulin promoter sequence in the genome of transgenic mice is probably not responsible *per se*, for decreasing insulin production through competition for factors necessary for insulin gene expression.

No lymphocytic infiltration

Lymphocytic infiltration was not observed in the pancreata of diabetic mice, or at any stage between days 2 and 37, regardless of haplotype. Occasional lymphocyte-like cells were sometimes seen but immunoperoxidase staining for the T cell surface molecules CD4 (L3T4) and CD8 (Lyt2) failed to identify these cells as T cells (not shown). This is in striking contrast to the marked infiltration seen in two other models of diabetes, namely the genetically IDDM-susceptible NOD mice and BB/W rats[9,10].

Diabetes in thymectomized mice

Our present findings suggested that diabetes did not result from an immune response either to self or foreign class I MHC molecules expressed in the β cells. The very early onset of the disease, the lack of lymphocytic infiltration and the occurrence of diabetes in mice genetically tolerant of the H-2K^b molecule, all tend to exclude a pathogenic role for T cells. We could not, however, dismiss the possibility that the H-2K^b glycoprotein expressed in the β cells of mice with the b haplotype was not identical to the endogenous H-2K^b expressed in the thymus and haemopoietic system. For example, a splicing variation of the transgene transcription product might generate a novel molecule[11]. Alternatively, the abundant class I heavy chain glycoprotein produced in the pancreas may not be associated with β_2 microglobulin and may appear in the membrane as a variant form[12]. To eliminate the possibility that a few T cells, in mice syngeneic with respect to the transgene, might have responded to an altered form of the class I molecules, neonatal thymectomy[13] was performed on 18 mice of the (C57BL/6×SJL)F_2 diabetic lineages. Some of these thymectomized mice were later grafted with skin from BALB/c donors (haplotype

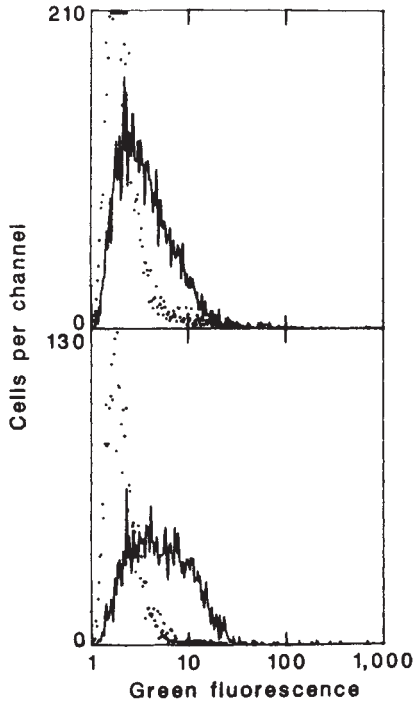


Fig. 4 Surface expression of H-2K^b on islet cells. Upper panel shows low level surface expression of H-2K^b on cells from non-transgenic mice (haplotype bs); lower panel shows increased level on cells from transgenic mice.
Methods. Islets were isolated from 14-day-old offspring of the 31-2 lineage by a modification of the collagenase digestion technique[29]. Prior to genotype identification, islets from individual mice were cultured free-floating at 37 °C in an atmosphere of 5% CO_2/95% air in plastic petri dishes containing 5 ml of RPMI-1640 plus 5% FCS. Following 5 days of culture, islets derived from transgenic or non-transgenic mice were pooled, washed three times in Ca^{2+}/Mg^{2+} free Hank's balanced salt solution containing 2 mM EDTA, trypsin-treated and dispersed into single cells by gentle aspiration and expulsion through a 1 ml plastic pipette. After 60 min recovery incubation at 37 °C, islet cells were stained with a monoclonal antibody against H-2K^b (B8-24-3, ref. 33) at 4 °C for 30 min, followed by washing and addition of fluorescein-conjugated anti-mouse Ig (Silenius). As a negative control, cells were stained with a monoclonal antibody against the H-2K^k protein (11-4-1, ref. 41). Approximately 5,000 cells per sample were analysed by flow cytometry in a fluorescence-activated cell sorter. The intensity of fluorescence (proportional to the level of H-2K^b proteins) is shown in arbitrary units on a three-decade logarithmic scale.

dd). They failed to reject the skin graft, which establishes their immune incompetence as clearly documented by previous work[14]. Six of the 18 mice were found to be transgenic and all six (four of bs haplotype, two ss) developed IDDM at the same age as nonthymectomized transgenic mice. All mice were killed at the end of the experiment to check for thymus remnants and none were found. These results argue strongly against a role for T cells in the pathogenesis of IDDM in class I transgenic mice.

Defect in insulin secretion

Immunoperoxidase staining and histology of pancreas tissue from pre-diabetic mice revealed the presence of insulin in apparently histologically normal islet tissue (Fig. 3, second and third panel). We wished to determine if these islets were in fact able to secrete the insulin they produced. Pancreatic islets were isolated from 14-day-old prediabetic transgenic mice as well as from matched controls and their capacity to secrete insulin was compared (Fig. 5). Chronic insulin release was significantly lower from islets isolated from transgenic mice than from islets

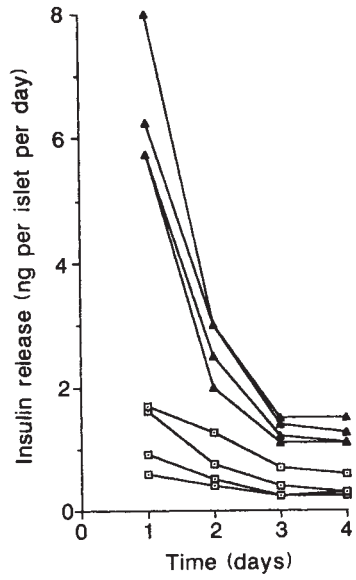


Fig. 5 Defective insulin secretion from pancreatic islets isolated from prediabetic transgenic mice. Chronic insulin release (insulin secretion) was significantly lower ($P<0.01$ for each point) from islets isolated from transgenic mice (□) than from nontransgenic littermates (▲). Each curve represents insulin secretion from islets isolated from a single mouse with each point being the mean of duplicate cultures.
Methods. Islets were isolated from the pancreas of 14-day-old offspring of the 31-2 lineage by a modification of the collagenase digestion technique[29]. Groups of 20 hand-picked islets were cultured free floating at 37 °C in an atmosphere of 5% CO_2/95% air, in 24-well multiwell plates (Linbro, Flow Laboratories) containing 1 ml of RPMI 1640 plus 10 mM glucose and 5% fetal calf serum. At 24 h intervals, medium samples (0.5 ml) were removed from each culture well, centrifuged and stored at −20 °C for subsequent assay of insulin by radioimmunoassay using rat insulin (Novo Biolabs, Denmark) as standard. Fresh medium (0.5 ml) was added to replenish each culture well.

of nontransgenic littermates. Thus, even at an age when clinical diabetes was not evident, the transgenic mice showed a defect in β cell function.

Discussion

We interpret the results described here as indicating that overexpression of class I MHC molecules in the β cells of the pancreatic islets is sufficient to induce β cell dysfunction and IDDM without T cell involvement. Four lines of evidence support this interpretation. First, mice with the b haplotype, genetically tolerant of the transgene product, developed diabetes when H-2K^b was expressed in their β cells. Second, the very early onset of diabetes suggests that the disease process occurred even before T cells had matured and emigrated from the thymus. Third, lymphocytic infiltration was not detected in the pancreas at any stage between 2 and 37 days. Finally, diabetes did occur in transgenic mice depleted of T lymphocytes by neonatal thymectomy. Although T cells are apparently not involved in the pathogenesis of diabetes in our model, in contrast to the spontaneous murine model of IDDM in the NOD mouse[9,15], the absence of a T cell infiltrate in the islets of transgenic mice, allogeneic with respect to the transgene, requires an explanation. There are at least four possibilities. First, the transgene product may escape T cell recognition because it is not expressed on the β cell membrane. This is not, however, supported by our immunofluorescence and flow cytometry studies (Fig. 4). Second, the transgene may be expressed as an altered form of the H-2K^b molecule, not seen by T cells unless associated with endogenous MHC molecules which are poorly expressed in the

β cells. Third, the transgene may fail to act as a restriction element for T cells by not associating with other normal endogenous cell surface molecules. Fourth, the transgene may be expressed as a classical class I MHC molecule during ontogeny and T cells are rendered tolerant to it. This last possibility implies either that induction of T cell tolerance is not a process confined to the intrathymic environment, or that it results from the secretion or shedding of molecules from β cells which reach the thymus where they are reprocessed to induce tolerance. The latter explanation was invoked in a previous report of tolerance in transgenic mice expressing, early in ontogeny, the SV40 T antigen under the control of the insulin promoter[16]. Whether or not T cells are tolerant of the nonself H-2K^b product in our model is being studied by classical cellular immunology techniques and the results will be reported elsewhere.

Our data show that in diabetic transgenics there is a defect in insulin secretion by β cells. How does the over expression of class I MHC molecules in the β cells result in such impairment? A characteristic of MHC class I molecules is their ability to bind peptides[17]. Since molecules synthesized within the cell may associate with class I glycoproteins[18], there may be some interaction between the abundant transgene product and either proinsulin or insulin itself. This could result in interference with the normal insulin processing and secretion pathway. An analogous interaction has been demonstrated between class I and an adenovirus glycoprotein in the endoplasmic reticulum[19]. Furthermore, non-covalent interactions have been demonstrated between class I molecules and membrane receptors involved in metabolism and growth[20-23]. We cannot, at this stage, dismiss the possibility that over-production of any protein would be associated with deleterious effects on β cell function. Previous studies on transgenic mice expressing, in the β cells, SV40 T antigen[16] and human insulin[24] did not describe any functional impairment as a result of the production of these ectopic proteins. In contrast, Sartvetnick *et al.*[25] and Lo *et al.*[26] have observed disruption of β cells through high level expression of class II MHC proteins, apparently with no lymphocytic involvement. These results, together with our present data on class I MHC gene expression, strongly support the hypothesis that upregulation of MHC gene products in β cells is sufficient *per se* to cause destruction of the cells.

It is possible that in the spontaneous models of IDDM, the BB/W rat and the NOD mouse, a primary, common destructive pathway is mediated through the upregulation and over-expression of Class I molecules triggered, for example, by cytokine production from inflammatory cells responding to a local virus infection, by virus infection directly, or by aberrant gene expression. In humans, Foulis *et al.*[27] found marked hyper-expression of class I molecules in all islet cells, and aberrant expression of class II molecules in residual β cells. They concluded that, within a given islet, hyperexpression of class I molecules preceded aberrant expression of class II and lymphocyte infiltration. Furthermore, over-expression of MHC molecules, in particular class I, is the common denominator when impairment of β cell function occurs in the absence of lymphocytic infiltration, as after infection of islets *in vitro* with the diabetogenic reovirus 1 or 3 (ref. 28) or exposure of islets *in vitro* to various cytokines[29-32]. These observations, together with the results of the present study using transgenic mice, lead us to postulate that the primary event leading to β cell destruction is the over-production of class I MHC molecules. It remains to be seen whether this effect is unique to pancreatic β cells or whether it may be a general phenomenon applicable to other models of autoimmune disease.

We thank Dr Alan Harris for critical reading of the manuscript and Ms Karen Holmberg, Ms Leonie Malcolm, Ms Audrey Szalai, Ms Anna Butt, Ms Leonie Oxbrow and Ms Angelina Iscaro for technical assistance. We are grateful to Dr D. Hanahan for the rat insulin promoter sequence and to Dr E. Weiss for the H-2K^b gene clone. The investigations were supported by grants from the National Health and Medical Research Council of Australia, the Utah Foundation, the Buckland Foundation, the Apex-Australian Diabetes Foundation Research and Education Grant, the Jack Brockhoff Foundation, the Sunshine and H. B. McKay Charitable Trust, the Kellion Diabetes Foundation and Hoechst Australia.

Received 8 March; accepted 27 April 1988

1. Dausset, J. & Hors, J. *Diabetes Res.* **1**, 115-123 (1984).
2. Drell, D. W. & Notkins, A. L. *Diabetologia* **30**, 132-143 (1987).
3. Sprent, J., Webb, S. R. *Adv. Immun.* **41**, 39-133 (1987).
4. Kappler, J. W., Roehm, M. & Marrack, P. *Cell.* **49**, 273-280 (1987).
5. Arnold, B. *et al. Proc. natn. Acad. Sci. U.S.A.* **85**, 2269-2273 (1988).
6. Baekkeskov, S. *et al. Proc. natn. Acad. Sci. U.S.A.* **78**, 6456-6460 (1981).
7. Hanahan, D. *Nature* **315**, 115-122 (1985).
8. Wilkie, T. M., Brinster, R. L. & Palmiter, R. D. *Devl. Biol.* **118**, 9-18 (1986).
9. Makino, S. *et al. Expl. Anim.* **29**, 1-14 (1980).
10. Seemayer, T. A., Tannenbaum, G. S., Goldman, H. & Colle, E. *Am. J. Path.* **106**, 237-249 (1982).
11. Lew, A. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 6084-6088 (1986).
12. Allen, H., Fraser, J., Flyer, D., Calvin S. & Flavell, R. *Proc. natn. Acad. Sci. U.S.A.* **83**, 7447-7451 (1986).
13. Miller, J. F. A. P. *Lancet* **ii**, 748-749 (1961).
14. Miller, J. F. A. P. *Proc. R. Soc.* **156**, 410-428 (1962).
15. Charlton, B. & Mandel, T. E. *Diabetes* (in the press).
16. Adams, T. E., Alpert, S. & Hanahan, D. *Nature* **325**, 223-228 (1987).
17. Bjorkman, P. J. *et al. Nature* **329**, 512-518 (1987).
18. Germain, R. N. *Nature* **322**, 687-689 (1986).
19. Andersson, M., Pääbo, S., Nilsson, T. & Peterson, P. *Cell* **43**, 215-222 (1985).
20. Schreiber, A. B., Schlessinger, J. & Edidin, M. *J. Cell Biol.* **98**, 725-731 (1984).
21. Kittur, D., Shimizu, Y., De Mars, R. & Edidin, M. *Proc. natn. Acad. Sci. U.S.A.* **84**, 1351-1355 (1987).
22. Due, C., Simonsen, M. & Olsson, L. *Proc. natn. Acad. Sci. U.S.A.* **83**, 6007-6011 (1986).
23. Phillips, M. L., Moule, M. L., Delovitch, T. L. & Yip, C. C. *Proc. natn. Acad. Sci. U.S.A.* **83**, 3474-3478 (1986).
24. Selden, R. F., Skoskiewicz, M. J., Howie, K. B., Russell, P. S. & Goodman, H. M. *Nature* **321**, 525-528 (1986).
25. Sarvetnick, N. *et al. Cell* **52**, 773-782 (1988).
26. Lo, D. *et al. Cell* **53**, 159-168 (1988).
27. Foulis, A. K., Farquharson, M. A. & Hardman, R. *Diabetologia* **30**, 333-343 (1987).
28. Campbell, I. L., Harrison, L. C., Ashcroft, P. G. & Jack, I. *Diabetes* **37**, 362-365 (1988).
29. Campbell, I. L., Wong, G. H. W., Schrader, J. W. & Harrison, L. C. *Diabetes* **34**, 1205-1209 (1985).
30. Campbell, I. L., Oxbrow, L., West, J. & Harrison, L. C. *Molec. Endocr.* **2**, 101-107 (1988).
31. Mandrup-Poulsen, T. *et al. Diabetes* **36**, 641-647 (1987).
32. Pukel, C., Baquerizo, H. & Rabinovitch, A. *Diabetes* **37**, 133-136 (1988).
33. Köhler, G., Fischer Lindahl, K. & Heusser, C. in *The Immune System* Vol. 2, 202-208 (Karger, Basel, 1981).
34. Weiss, E. H. *et al. Nature* **310**, 650-655 (1984).
35. Marsh, J. L., Erfle, M. & Wykes, E. J. *Gene* **32**, 481-485 (1984).
36. Weiss, E. H. *et al. EMBO J* **2**, 453-462 (1983).
37. Mellor, A. L. *et al. Nature* **298**, 529-534 (1982).
38. Kimura, A., Israel, A., Le Bail, O. & Kourilsky, P. *Cell* **44**, 261-272 (1986).
39. Hogan, B., Costantini, F. & Lacy, E. in *Manipulating the Mouse Embryo. A laboratory Manual* (Cold Spring Harbor Laboratory, New York, 1986).
40. Sternberger, L. A. in *Immunocytochemistry* 2nd edn. (Wiley, New York, 1979).
41. Oi, V. T., Jones, P. P., Goding, J. W., Herzenberg, L. A. & Herzenberg, L. A. *Curr. Topics Microbiol. Immun.* **81**, 115-129 (1978).

The theory of gamma-ray emergence in supernova 1987A

Philip A. Pinto & S. E. Woosley

Lick Observatory and Board of Studies in Astronomy and Astrophysics, University of California, Santa Cruz, California 95064, USA

Since June 1987 we have had little cause to doubt the presence of radioactivity in supernova 1987A. The optical light curve has tracked the 77.1-day half-life of ^{56}Co to better than 1% (refs 1–3), at least through November, demonstrating conclusively the synthesis of $0.075M_{\odot}$ of radioactive ^{56}Ni in the explosion[4]. It was anticipated that the decay of ^{56}Co to ^{56}Fe would give rise to detectable γ-ray line emission[5,6] at 847 and 1,238 keV with a peak flux[7] of 10^{-3} photons cm^{-2} s^{-1} about one year after the explosion. Many calculations of the light curves for these lines were made for the particular case of SN1987A (refs 8–14), and in August, both lines were detected[15–18], with a strength within a factor of two of 10^{-3} photons cm^{-2} s^{-1}, but about six months earlier than predicted (see ref. 8 for example). Here we show that the early emergence of γ-rays can be accounted for in a 'mixed' model, in which an approximately isotropic process destroys chemical segregation with respect to radial mass coordinate and velocity.

The emergence in August 1987 of a hard X-ray continuum[19,20] showed that the Comptonized tails of γ-rays from the decay of ^{56}Co were beginning to break through[8,20–25], and an infrared spectrum taken in October[26] displayed a prominent line at 10.5 μm, presumably due to ^{56}Co in the core. These observations agree well with theoretical expectations that radioactive nuclei, particularly ^{56}Ni, should be produced in supernovae[27–30] in amounts of the order of $0.1M_{\odot}$[31,32]. Because the amount of ^{56}Co produced in the explosion is well known from the optical light curve, the early appearance of γ-ray lines implies that at least a trace of ^{56}Co lies at a smaller column depth than would follow from a strictly spherically symmetric model which agrees with the light curve. Because also the ^{56}Co must have been made (as ^{56}Ni) only in the bottom-most layers ejected by the supernova, and because a more rapid but still spherically symmetric expansion than assumed in the model would violate other observational constraints[4], the early appearance of γ-ray lines implies that spherical symmetry in the expansion must have been broken. This could have occurred through any combination of mixing, clumping or the formation of jets.

The blast wave responsible for producing ^{56}Ni in what had been the silicon shell of the presupernova star was itself nearly spherically symmetric, but as the freshly-synthesized heavy elements, with velocities 2,000 to 3,000 km s^{-1}, ran into the low density hydrogen envelope, instabilities caused non-uniform deceleration of the core material[33]. In the end not all of the ^{56}Ni was slowed to below 1,000 km s^{-1} as predicted in the spherically symmetric models. Also, following the explosion the decay of ^{56}Ni and ^{56}Co deposited about 9×10^{17} erg g^{-1} in the expanding material. Without the tamping effect of the overlying ejecta, this energy would have accelerated the ashes of the decay to a speed of about 4,000 km s^{-1}. In a one-dimensional model[4], however, the speed of the ^{56}Co stays less than 1,200 km s^{-1} because its mass is so much smaller than that of the overlying oxygen and helium. In general, it is probably true that the bulk of the ^{56}Co comes out slowly but, by Rayleigh–Taylor instability, a small fraction may achieve speeds up to 4,000 km s^{-1} and thus spread out not only through the helium core but even into the hydrogen envelope[4]. This mixing could take the form of finger-like protuberances, but here for computational ease is assumed isotropic. Mixing has been previously invoked by a number of authors both to explain the early appearance of the hard X-rays

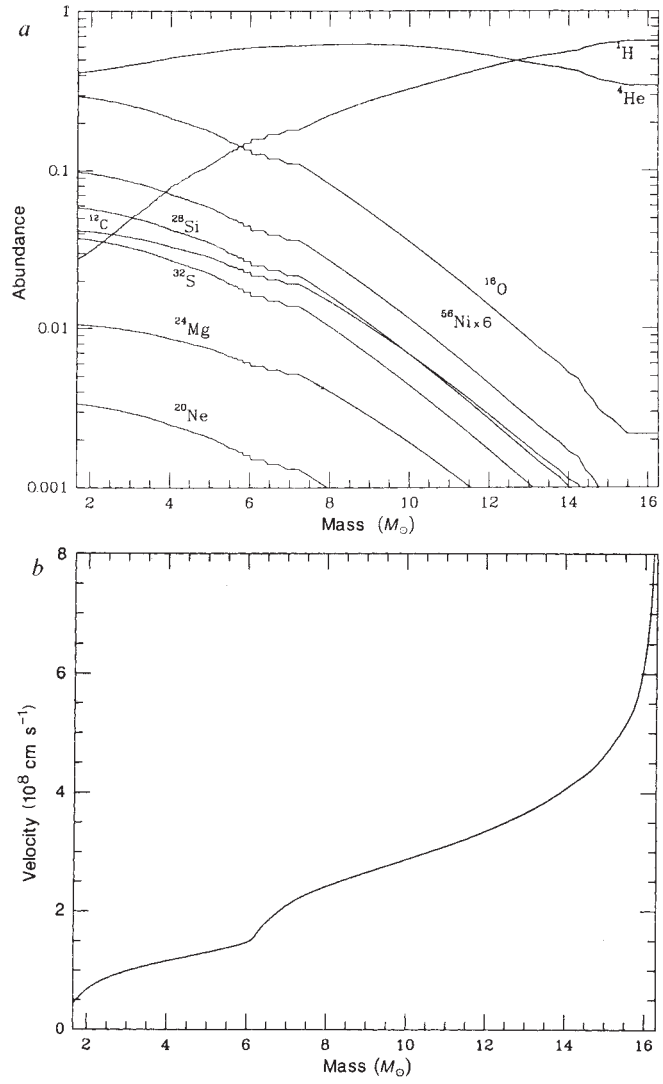


Fig. 1 Composition (*a*) and velocity (*b*) of the mixed Model 10HMM as a function of mass in the ejecta. The ^{56}Ni mass fraction (which turns into ^{56}Co early on) has been multiplied by a factor of six for easy display.

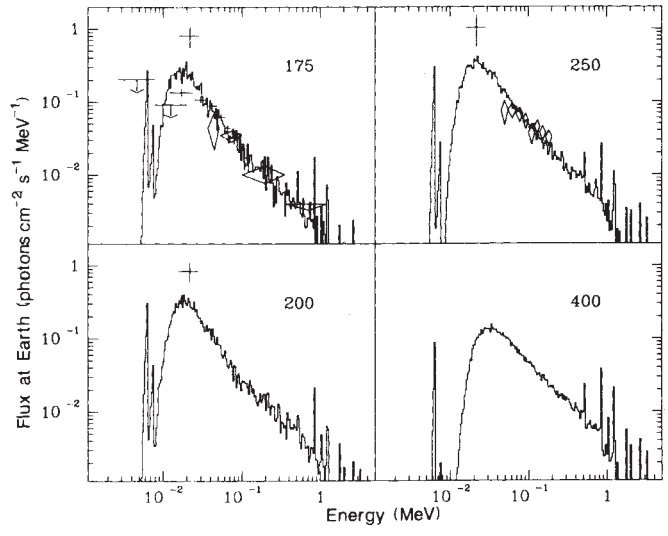
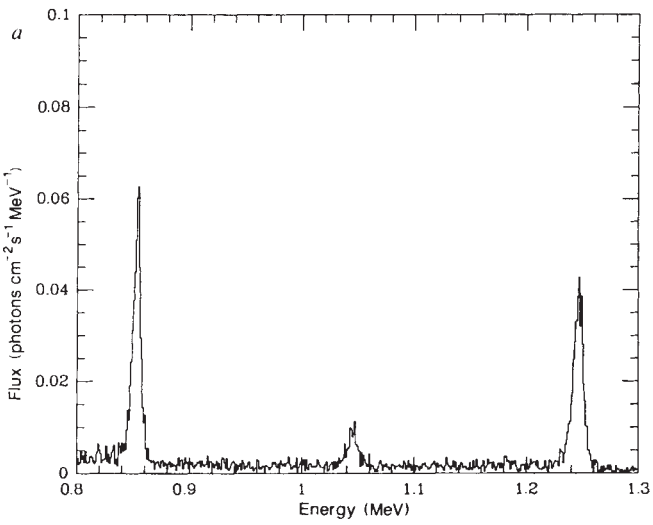
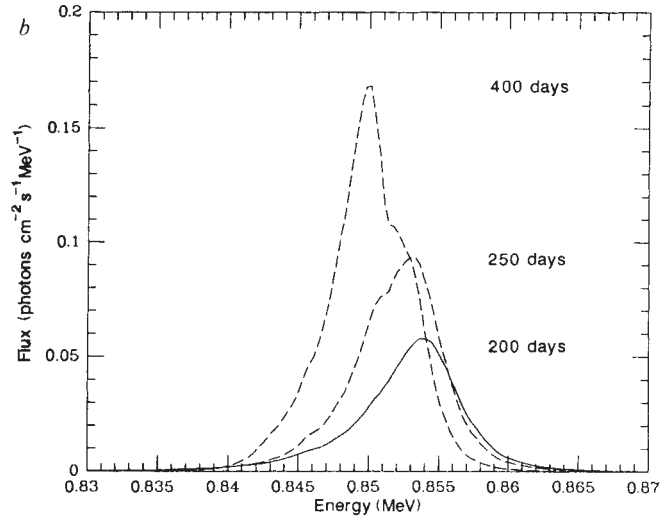


Fig. 2 Hard X-ray spectra on days 175, 200, 250, and 400 of the supernova, obtained by a Monte Carlo calculation based upon Model 10HMM. Data on day 175 are from ref. 20. Data from day 250 are from ref. 14 and R. Wilson (personal communication).





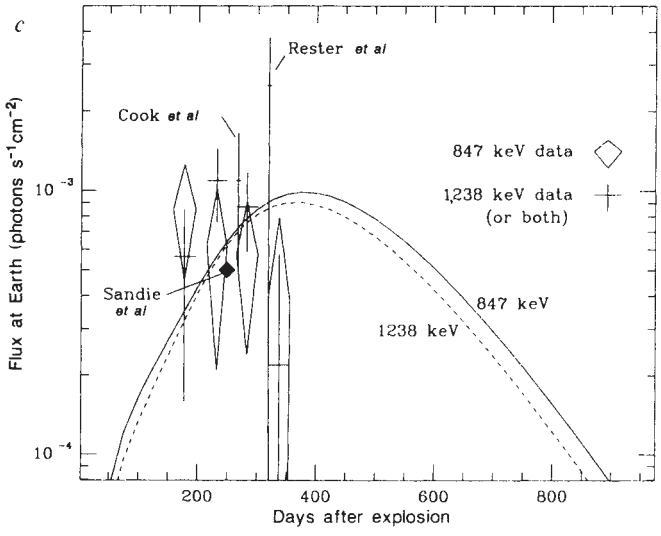


Fig. 3 Gamma-ray lines. *a*, Line profiles at day 200 for the 847 and 1,238 keV lines obtained by Monte Carlo calculation. The line strength in a given energy interval is the area under this curve. *b*, γ-ray spectrum in the 830 to 870 keV region including only those photons that escape without interaction. As the supernova ages the lines become progressively narrower and redder as deeper material having lower velocity is uncovered. *c*, The light curves of these unscattered photons in both 847 and 1,238 keV lines for Model 10HMM compared to observational data[15-17]. At late times the ratio of the 847 and 1,238 keV lines approaches its laboratory value (0.676). The dotted line in the upper right is the flux of the 847 keV line in the case of zero optical depth. The data from SMM (ref. 15) show marginal evidence for a decreasing signal when our model predicts that it should be slowly increasing.

in August and September[8,14,20,23-25] and the smooth nature of the optical light curve[4,23,34-36]. There also exists evidence in the Cas-A supernova remnant for mixing having occurred sometime after the explosion. In particular, some of the 'fast moving knots' that have a composition typical of explosive oxygen burning (large abundances of intermediate mass elements like sulphur and argon) are found to be moving faster than other knots whose composition shows that they reached too low a temperature during the explosion to burn oxygen (R. P. Kirshner, private communication).

For the time being, we do not explore the implications of 'clumpy' models or of jets. A high velocity jet (as opposed to the 'fingers' described above) is an unlikely explanation for the early appearance of the X-rays and γ-rays because observations prior to August did not detect a strong flux of either. Clumping, on the other hand, probably has occurred to some extent. Again note the appearance of the Cas-A supernova remnant. Observationally the γ-line spectrum of a clumped model would differ from a mixed model because the ^{56}Co and other heavy elements could still be confined to a small region of low velocity. If observations over the next few months show that no portion of the ^{56}Co is moving with speeds $\gtrsim$1,500 km s^{-1}, then we shall have to consider clumped models which, unfortunately, are inherently more complicated.

The mixed model we shall consider is Model 10HMM (Fig. 1), which is based upon Model 10HM[4,34,35], but has been mixed just a little more. Because of the exponential sensitivity of the γ-line flux to column depth, the observations are sensitive to even a very small amount of ^{56}Co moving at high velocity (Fig. 1). For now the mixing operation, though meant to simulate a real physical phenomenon, is carried out in an arbitrary fashion. Model 10HMM (and 10HM) both follow the explosion (1.3×10^{51} erg) of a 6-$M_\odot$ helium core capped by a blue supergiant envelope of $10M_\odot$. The explosion, simulated by a piston at the base of the silicon shell, ejects all overlying material and explosively processes a portion of the silicon shell into ^{56}Ni. To agree with observations of the tail of the light curve this ^{56}Ni mass was artificially adjusted to be $0.075M_\odot$. The light curve and photospheric velocity and temperature histories of this model agree well with what has been observed for SN1987A and the small amount of extra mixing required to obtain Model 10HMM here from Model 10HM does greatly not alter the light curve or photospheric history.

The γ-line profiles and light curves were calculated for this model using a combination of impact parameter integration and Monte Carlo techniques[8]. For the first time in any work that also included the necessary compositional mixing, adequate energy resolution was included in the Monte Carlo calculation to resolve both the narrow γ-lines and their surrounding Compton continuum. The results for the hard X-ray spectrum on days 175, 200, 250 and 400 are shown in Fig. 2. Also shown are a magnification of the flux spectrum in an energy range around 847 and 1,238 keV (Fig. 3*a* and 3*b*) and the expected light curves (Fig. 3*c*) in the 847 keV and 1,238 keV lines of ^{56}Co decay. Although Fig. 3*a* shows the Monte Carlo results and thus includes a continuum of scattered photons, Fig. 3*b* and 3*c*

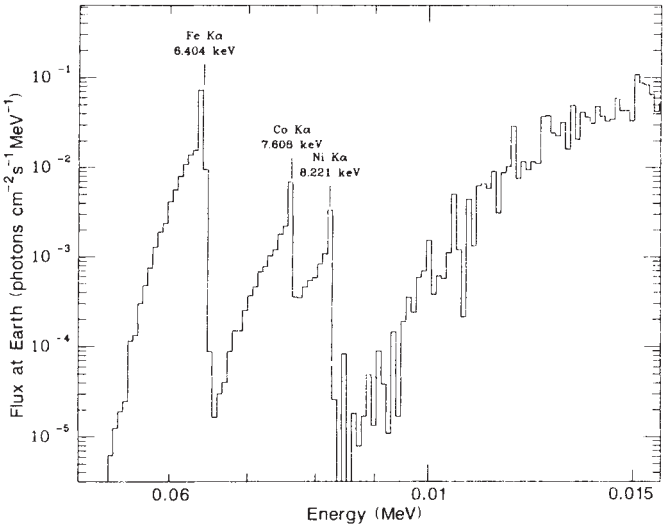


Fig. 4 The hard X-ray spectrum of Model 10HMM at day 250, showing the fluorescence Kα lines of nickel, cobalt and iron.

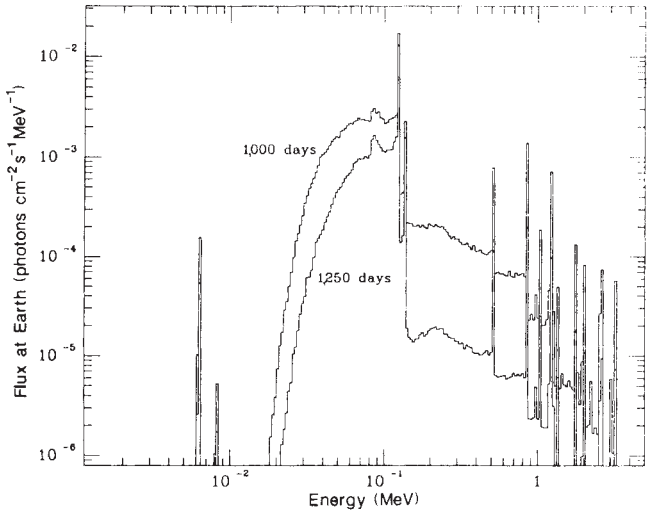


Fig. 5 Gamma-ray spectrum of Model 10HMM on days 1,000 and 1,250. Most of the energy is produced by the decay of ^{57}Co. Note the strong narrow line at 122 keV. The fluorescence lines of iron and cobalt have become very weak.

were calculated using an impact parameter code and are only for photons that escape without interaction and hence form the line 'core'. Any real detector would also see additional flux from the continuum, in an amount depending upon the device's energy resolution and the manner in which the background is subtracted.

Similarly, if the background count rate is a substantial fraction of the line strength at centre only photons in a narrow energy interval about the centroid may be inferred actually to constitute the signal. Thus photons in the line wings, which are substantial, may be overlooked. The importance of these two effects becomes apparent when we numerically integrate the flux found in the Monte Carlo calculations (Fig. 3*a*) in various energy windows. The results, given in Table 1, show that a sampling of the spectrum around 847 keV at days 200 and 250 of the supernova including either a window of ~55 keV (ref. 15) or 81 keV (ref. 17) gives approximately twice the flux that would be sampled by a germanium detector in an 8-keV wide band. This is due both to the extended 'wings' on the lines and substantial jump in the Compton continuum on the red side of the line. Here the narrow line at 847 keV has a width of about 6 keV (full width half maximum) and a centroid indicated by the range of energies given in Table 1. Generally speaking the fluxes given in Table 1 are consistent with all other γ-ray observations to date (each of which have large associated error bars) and are also consistent with the hard X-ray flux reported[20] on day 175 (Figs 2 and 3*c*).

Table 1 Model 10HMM gamma-'line' fluxes

Energy range (MeV)	day 200	day 250	day 400
0.806–0.888	7.1×10^{-4}	1.1×10^{-3}	1.3×10^{-3}
0.825–0.881	6.4×10^{-4}	9.4×10^{-4}	1.2×10^{-3}
0.848–0.856	3.3×10^{-4}	—	—
0.847–0.855	—	5.4×10^{-4}	—
0.846–0.853	—	—	6.8×10^{-4}
1.191–1.285	5.6×10^{-4}	8.2×10^{-4}	9.8×10^{-4}
1.240–1.250	3.0×10^{-4}	—	—
1.239–1.249	—	4.7×10^{-4}	—
1.236–1.246	—	—	5.8×10^{-4}

Also apparent in the spectrum (Fig. 4) are intense 'soft X-ray' emission lines that come from K-shell emission of iron, nickel, and (radioactive) cobalt synthesized during the explosion and mixed into the hydrogen envelope. Lines of other heavy elements should be present at a reduced level. Table 2 gives the calculated fluxes and centroids for the two strongest lines at several times. These lines have much smaller blue shifts than the γ-ray lines because they are the result of ionization at small optical depth followed by isotropic emission. Most of the flux thus comes from larger impact parameters where the projected velocity along the line of sight is lower. The γ-lines on the other hand, being essentially unscattered, come from cobalt at smaller radii but, subtending a smaller line of sight to the centre of the supernova, display larger blue shifts.

Though observations and calculations thus far have, for the most part, been concerned solely with the high-energy radiation emitted during the decay of ^{56}Co, at later times (a few years) the radiation from ^{57}Co decaying to ^{57}Fe will become important and potentially observable[10,11,37,38]. Figure 5 shows the expected spectrum on days 1,000 and 1,250 assuming that ^{57}Co (made as ^{57}Ni) exists in a ratio to ^{56}Co (or ^{56}Ni) like the solar ratio of ^{57}Fe/^{56}Fe, 2.4×10^{-2}. Note especially the distinct signature of ^{57}Co at 122 keV and 136 keV, still with a significant Compton tail. The observation of ^{57}Co decay lines will provide additional unique information regarding the nucleosynthesis of an isotope sensitive to the neutron excess and freeze out conditions in the deepest layers to be ejected. Its discovery should be given high priority by future missions, such as the Gamma-Ray Observatory. One thousand days after explosion, the 122 keV line will present a flux at Earth of $\sim8\times10^{-5}$ photons cm^{-2} s^{-1}, whereas the iron Kα flux will have dropped to $\sim1\times10^{-8}$ photons cm^{-2} s^{-1}.

We conclude by emphasizing the importance of immediate accurate measurements of the line centroid energy and shapes of the decay lines of ^{56}Co. In addition to photometry these are the primary scientific goals of γ-line astronomy of SN 1987A, as they will give unique and unambiguous evidence for or against the substantial amount of outward mixing we and others have invoked in the models. As time passes, however, the signal will

Table 2 X-ray line strengths (10^{-4} photons cm^{-2} s^{-1})

Element energy (keV)	day 200 flux	day 250 flux	day 400 flux
Fe			
6.404	8.7	8.2	2.3
Co			
7.609	1.6	0.97	0.056

become increasingly dominated by the large fraction of cobalt moving at slower speeds (Figs 1 and 3*b*) and precious information will have been irretrievably lost. While similar information can also be obtained from analysis of the Co II infrared line near 10.5 μm (ref. 26), the γ-ray lines, for the time being anyway, originate from further out in the supernova than the infrared. Thus their line centroids are a more sensitive measure of mixing. This novel aspect of the greater penetrating power of infrared photons is a consequence of an opacity determined by atomic level populations rather than total electron abundance, but the advantage will soon be lost. We urge that both the γ-lines and the 10.5-μm infrared line be continually monitored at high-energy resolution.

This research was supported by the National Science Foundation (AST 84-18185) and by the NASA Theory Program. We are grateful for the allocation of a block of time at the San Diego Supercomputer Center that allowed us to make the calculations reported here in a timely fashion. Liek Observatory Contribution No. 453.

Received 8 February; accepted 12 April 1988.

1. Menzies, J. W. *et al. Mon. Not. R. astr. Soc.* **227**, 39 (1987).
2. Catchpole, R. M. *et al. Mon. Not. R. astr. Soc.* **229**, 15 (1987).
3. Hamuy, M., Suntzeff, N. B., Gonzalez, R. & Martin, G. *Astr. J.* **95**, 63-83 (1988).
4. Woosley, S. E. *Astrophys. J.* (in the press).
5. Clayton, D. D., Colgate, S. A. & Fishman, G. J. *Astrophys. J.* **155**, 75-82 (1969).
6. Clayton, D. D. in *Essays in Nuclear Astrophysics* (eds Barnes, C. A., Clayton, D. D. & Schramm, D. N.) 401-426 (Cambridge University Press, 1982).
7. Woosley, S. E., Axelrod, T. S. & Weaver, T. A. *Comments Nucl. Part. Sci.* **9**, 185-197 (1981).
8. Pinto, P. A. & Woosley, S. E. *Astrophys. J.* (in the press).
9. Woosley, S. E., Pinto, P. A. & Ensman, L. *Astrophys. J.* **324**, 466-489 (1988).
10. Woosley, S. E. & Pinto, P. A. in *Proc. Workshop on Gamma-ray Spectroscopy of Astrophysical Sources* (eds Share, G. & Gehrels, N.) (American Institute of Physics, New York, in the press).
11. Chan, K. W. & Lingenfelter, R. E. *Astrophys. J.* **318**, L51-L56 (1987).
12. Gehrels, N., MacCullum, C. J. & Leventhal, M. *Astrophys. J.* **320**, L19-L?? (1987).
13. Shibazaki, N. & Ebisuzaki, T. *Astrophys. J.* **327**, L5-L8 (1988).
14. Ebisuzaki, T. & Shibazaki, N. *Astrophys. J.* **327**, L9-L12 (1988).
15. Matz, S. M., Share, G. H., Leising, M. D., Chupp, E. L. & Vestrand, W. T. *Nature* **331**, 416-418 (1988).
16. Wilson, R. *et al. IAU Circ.* No. 4535 (1988).
17. Cook, W. R. *et al. IAU Circ.* No. 4527 (1988).
18. Rester, A. C. *et al. IAU Circ.* No. 4535 (1988).
19. Dotani, T. *et al. Nature* **330**, 230-231 (1987).
20. Sunyaev, R. *et al. Nature* **330**, 227-229 (1987).
21. McCray, R., Shull, J. M. & Sutherland, P. G. *Astrophys. J.* **317**, L73-L87 (1987).
22. Xu, Y., Sutherland, P. G., McCray, R. & Ross, R. R. *Astrophys. J.* (in the press).
23. Itoh, M., Kumugai, S., Shigeyama, T., Nomoto, K. & Nishimura, J. *Nature* **330**, 233-235 (1987).
24. Grebenev, S. & Sunyaev, R. *Sov. Astr. Lett.* **13**, 945 (1987).
25. Shull, J. M. & Xu. Y. in *Proc. George Mason Workshop in SN1987A* (ed. Kazanas, D.) (Cambridge University, in the press).
26. Rank, D. M. *et al. Nature* **331**, 505-506 (1988).
27. Suess, H. E. & Urey, H. C. *Rev. mod. Phys.* **28**, 53-74 (1956).
28. Pankey, T. thesis (Howard University, 1962).
29. Colgate, S. A. & McKee, C. *Astrophys. J.* **157**, 623-643 (1969).
30. Bodansky, D., Clayton, D. D. & Fowler, W. A. *Astrophys. J. Suppl.* **16**, 299-371 (1968).
31. Weaver, T. A. & Woosley, S. E. *Ann. N.Y. Acad. Sci.* **336**, 335-357 (1980).
32. Uomoto, A. & Kirshner, R. P. *Astrophys. J.* **308**, 685-690 (1986).
33. Chevalier, R. & Klein, R. I. *Astrophys. J.* **219**, 994-1007 (1978).
34. Woosley, S. E. in *Proc. IAU Colloq. 108: Atmospheric Diagnostics of Stellar Evolution* (ed. Nomoto, K.) (Reidel, Dordrecht, in the press).
35. Woosley, S. E. in *Proc. George Mason Workshop on SN1987A* (ed. Kazanas, D.) (Cambridge University, in the press).
36. Arnett, W. D. in *Proc. George Mason Workshop on SN1987A* (ed. Kazanas, D.) (Cambridge University, in the press).
37. Clayton, D. D. *Astrophys. J.* **188**, 155-157 (1974).
38. Woosley, S. E., Pinto, P. A., Martin, P. & Weaver, T. A. *Astrophys. J.* **318**, 664-673 (1987).

Unusual radio source MG1131+0456: a possible Einstein ring

J. N. Hewitt*, E. L. Turner†, D. P. Schneider‡, B. F. Burke , G. I. Langston‖# & C. R. Lawrence¶

*Haystack Observatory, NEROC, Westford, Massachusetts 01886, USA
†Princeton University Observatory, Peyton Hall, Princeton, New Jersey 08544, USA
‡School of Natural Sciences, Institute for Advanced Study, Princeton, New Jersey 08540, USA
‖Department of Physics and Research Laboratory for Electronics, Room 26-331, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
¶Mail Stop 105-24, California Institute of Technology, Pasadena, California 91125, USA

We report the discovery of extremely unusual structure in the radio source MG1131+0456. In a radio map with sub-arcsecond resolution the object appears as an elliptical ring of emission accompanied by a pair of more compact sources, nearly diametrically opposed and offset ~0.3 arc s to the south-west of the ring. There is a faint, slightly extended, optical counterpart. In its spectrum we detect a continuum but no emission lines; therefore the redshift of the counterpart is unknown. The radio morphology suggests that this source may be an example of an 'Einstein ring'[1], a highly symmetric case of gravitational lensing in which the source is imaged into a ring. Other explanations that we consider are excluded by the data. Another possibility is that MG1131+0456 represents a new type of (unlensed) astronomical object.

The radio source MG1131+0456 (probably 4C05.51) was first mapped by us with the National Radio Astronomy Observatory's Very Large Array (VLA) telescope as part of a project to study a large number of sources detected in the MIT-Green Bank (MG) survey[2]. Our original observation consisted of <2 min of integration at 5 GHz frequency with the A configuration of the VLA, but the resulting 'snapshot' map was of sufficient quality to reveal unusual ring-like structure in the source. The striking morphology of MG1131+0456 prompted us to carry out multi-frequency observations of the source during a 9-h period with the VLA on 20 September 1987; a full report of this extensive set of data will be published elsewhere. Figure 1 presents a total intensity map constructed from the 15-GHz data acquired on that date. To our knowledge, the radio morphology displayed in Fig. 1 is unprecedented, consisting of a marginally resolved,

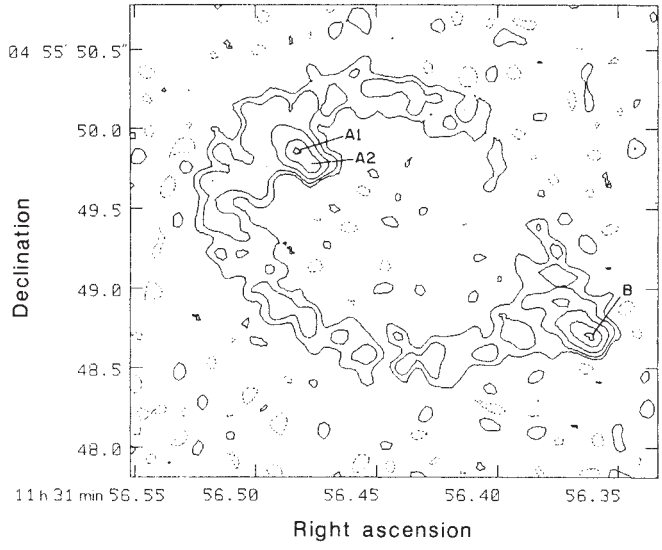


Fig. 1 Contour plot of 15 GHz radio image of MG1131+0456. The image was constructed from data acquired with the VLA in its A configuration during a 9-h observing run. The beam is well represented by a circular gaussian profile with FWHM (full width at half maximum) of 0.12″, and the noise level (1σ) is 115 μJy per beam. The contours are at −8%, 8%, 16%, 32%, 64% and 95% of the peak brightness of 2.6 mJy per beam.

Present address: Max-Planck-Institut für Radioastronomie, Auf dem Hugel 69, D-5300 Bonn 1, FRG.

elliptical ring of low surface brightness accompanied by two more compact sources. The brightness distribution in the map, excluding the compact sources, is well described by an ellipse with major and minor axes of 2.2″ and 1.6″, position angle of 70°, and centre at right ascension 11 h 31 min 56.44 s and declination +04° 55′ 49.4″ (J2000). Both the inner and outer edges of the ring are clearly defined. Each of the pair of sources has structure; the northern component is double (components A1 and A2), and the structure of the southern component (component B) is consistent with its being a close double or resolved (see Table 1). In addition to the elliptical ring and the two more compact components, a low-surface-brightness source (component C), to the south-west of the ring, is evident in a contour plot of a 5-GHz radio map (Fig. 2), also constructed from data acquired with the VLA in September 1987. Figure 3 shows plots of a row in each radio image that demonstrate that there is nearly no emission at the centre of the ring. The total flux density in the 15-GHz map, in the region of the interior of the ring more than three beamwidths from the ring, is $\sim 80 \pm 20$ μJy over an area of 0.6 arc s^2. In contrast, the surface brightness of the ring is ~25 mJy arc s^{-2}. The strength and position angle of linearly polarized emission at 5 GHz are indicated in Fig. 2. At this frequency, the compact components are not cleanly resolved from the ring; however, at both their positions the polarization is in the north-south direction. All other parts of the source appear to be polarized at angles that vary quite smoothly around the ring. Because we have detected polarized emission in the ring at only one frequency, we are unable to determine whether Faraday rotation has affected the position angle of polarization. Differential Faraday rotation across the source could affect the relative position angles at the positions of the compact components.

On 27 March 1987 we acquired an R band optical image of MG1131+0456 with the National Optical Astronomy Observatories' 4-m telescope on Kitt Peak equipped with the prime focus T12 CCD (charge-coupled device) camera. We identify a faint ($m_R \approx 22$) optical object near the position of the radio source; a finding chart is presented in Fig. 4. In seeing of 1.1″ the optical object was resolved; a fit of an elliptical gaussian profile to the image gives major and minor axes of 2.5″ and 2.1″, and a position angle of ~45°. Spectroscopic observations of MG1131+0456 on 1-2 April 1987 were made with the Cryogenic Camera and the R-C spectrograph on the 4-m telescope. The wavelength coverage was 4,500-8,900 Å with a resolution of 15 Å sampled at 4.3 Å per pixel. We acquired a total of 14,400 s of spectral data, but the observations were affected by bad weather and the data are not of high quality. We detected a red continuum, but no emission lines; therefore, we are as yet unable to measure the redshift of the optical counterpart. MG1131+0456 is not a known infrared source; it is listed in neither the IRAS (infrared astronomical satellite) point source catalogue, nor the small extended source catalogue[3,4]. Its high galactic latitude ($b^{II} = 60°$) indicates that it is likely to be extragalactic.

Ring-like morphology in radio sources is often seen in supernova renmants, H II regions, and planetary nebulae, but the optical and infrared properties of MG1131+0456 appear to be inconsistent with these interpretations. The radio spectral index

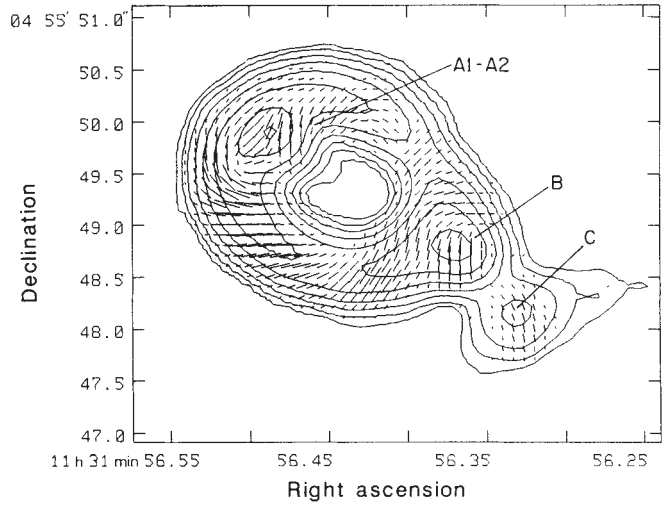


Fig. 2 Contour plot of 5 GHz radio image of MG1131+0456, constructed from data acquired during the same observing run as the image of Fig. 1. The beam has FWHM of 0.39″; the noise level (1σ) in the map is 50 μJy per beam. The contours are at 1%, 2%, 4%, 8%, 16%, 32%, 64% and 95% of the peak brightness of 23 mJy per beam. Superimposed on the contours are line segments representing the strength and position angle of linearly polarized emission at 5 GHz.

of 0.9 ± 0.02 (calculated from published flux densities at 1.4 and 5 GHz[2,5]) indicates that the emission mechanism is non-thermal and, along with the optical and infrared properties, is consistent with a radio galaxy. The morphology of MG1131+0456 strongly suggests that it is an instance of gravitational lensing. An object lined up with a foreground lens of circular symmetry is imaged into a ring, as Einstein noted[1]. A second object, displaced slightly from the centre of the lens, gives a principal image just outside the ring and a second image just inside the ring diametrically opposed to the first image. In this case a model can be constructed of a radio object imaged by a lens that is characterized by a simple and plausible gravitational potential. In such a model, the intrinsic structure of the MG1131+0456 would not be so unusual: a compact (possibly double) core accompanied by two radio lobes. One lobe of the radio source would fall directly behind the lens and be imaged into the ring; the compact core would fall near enough to the lens to be imaged twice (into A1-A2 and B); and the second lobe would fall far enough from the lens that only one bright image would be visible (C). The measured polarization angles of the compact sources support this interpretation, although the effects of Faraday rotation must be checked. The ellipticity of the ring implies that a mass distribution described by at least the monopole and quadrupole moments is required. We have calculated lensing models by the techniques of Schramm and Kayser[6] that incorporate the elliptical potential described by Blandford and Kochanek[7]. This potential is parameterized by its depth, which can be represented by a velocity dispersion σ_v, its ellipticity ε, its core radius s and

Table 1 Summary of radio data

Component	Frequency (GHz)	Right ascension (J2000)	Declination (J2000)	Peak brightness (mJy per beam)	Position angle with respect to centre of ellipse
A1	15	11 h 31 min 56.483 s	+04° 55′ 49.86″	2.6	+54°
A2	15	11 h 31 min 56.477 s	+04° 55′ 49.78″	2.5	+55°
B	15	11 h 31 min 56.362 s	+04° 55′ 48.70″	2.6	−121°
C	5	11 h 31 min 56.33 s	+04° 55′ 48.2″	2.2	−127°

The relative uncertainty between the positions measured at 15 GHz is ~0.01″. The integrated flux density of component C at 5 GHz is ~5 mJy. The position angles are defined in the usual astronomical sense so that north is at 0° (at the top in the figures) and east is at 90° (to the left in the figures).

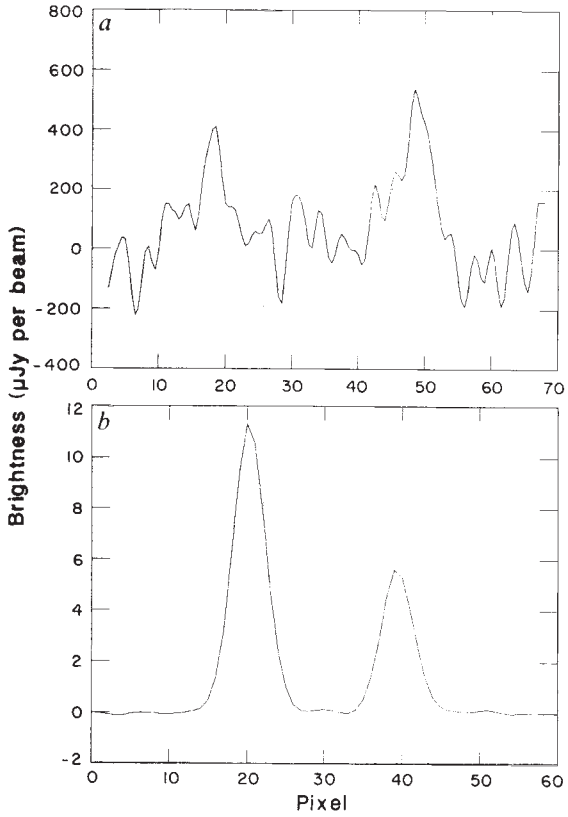


Fig. 3 *a*, Plot of a row (declination 04° 55′ 49.04″) in the 15 GHz map. One pixel is 0.03″. *b*, Plot of a row (declination 04° 55′ 49.31″) in the 5 GHz map. One pixel is 0.1″.

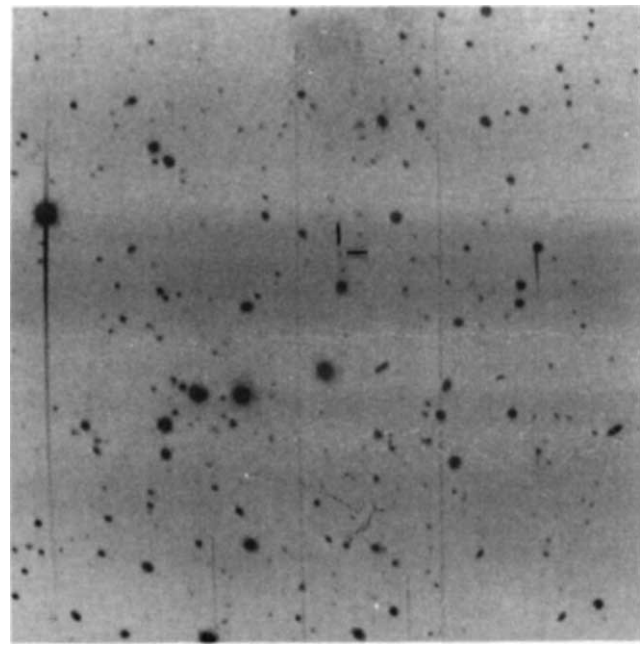


Fig. 4 R band optical image (4′ by 4′) of the MG1131+0456 field. North is at the top; east to the left. The counterpart to MG1131+0456 is indicated.

its hardness α (for $\varepsilon = 0$, $\alpha = \frac{1}{2}$ corresponds to an isothermal sphere lens and $\alpha = 0$ corresponds to a point mass lens). Assuming $q_0 = 0$ and arbitrary but reasonable lens and source redshifts of 0.5 and 1.0, we are able to reproduce the size and ellipticity of the ring with the choice of parameters $\sigma_v = 290$ km s^{-1}, $\varepsilon = 0.31$, $s = 100$ pc and $\alpha = \frac{1}{2}$. In this model, the ring is made up of four merging images of the portion of the source directly behind the lens. The approximate positions of the large (to the northwest) and small (to the south) gaps in the ring can be reproduced by moving the source a small distance from the centre of the lens. The relation of A1-A2 to B within the context of the lensing interpretation cannot be determined from available data; A1-A2 could be second and third images of B, or A1-A2 could be a single image of an intrinsically double B. The first case is difficult to reconcile with the small core radius required to suppress the central image of the ring. The second case would require a very 'soft' potential to account for the relatively large A1-A2 separation; in the circularly symmetric case the geometry would imply that $\alpha > \frac{1}{2}$. The position angles of A1-A2 and B with respect to the position angle of the ring are −16° and +169°, indicating that the compact component of the lensed source does not fall exactly on the major axis of the lens. Our simulations of lensing by elliptical potentials indicate that, for a doubly imaged source, the image outside the ring is rotated through a smaller angle than the image inside the ring, consistent with the geometry of MG1131+0456.

We have calculated the above solution merely to demonstrate that gravitational lensing is a possible mechanism for producing an ellipse; modelling that incorporates all the available data should be done to further test the viability of the lensing interpretation of the data. One may well object that the formation of an Einstein ring requires an extremely improbable alignment of a source and a circularly symmetric lens. But perfect alignment is not required if the source is extended. The requirement of circular symmetry is also relaxed because the multiple images (four images for an elliptical potential) merge, giving the appearance of a ring. Our numerical simulations of lensing by an elliptical ($\varepsilon = 0.3$) potential at a redshift of 0.5 indicate that, for a source at a redshift of 1.0, a source radius of only 0.3″ is required (we assume a cosmological deceleration parameter $q_0 = 0$). We can compare this lensing situation to that of the formation of double images by an isothermal sphere lens parameterized by a velocity dispersion typical of galaxies, 200 km s^{-1}. For the same source and lens redshifts, the source must lie within 0.45″ of the lens to be multiply imaged[8]. Therefore, for the probability of producing an elliptical ring to be equal to the probability of multiple imaging by a galaxy-sized isothermal sphere, a source radius of ~0.8″ is needed. This is comparable to the size of the lobes of extragalactic radio sources. Given the size of our sample of sources mapped with the VLA (several thousand, of which approximately half have extended components), it would not be surprising to observe several cases of ring-like objects.

The properties of MG1131+0456 strongly suggest gravitational lensing of an extended source by an elliptical potential. Other known types of astrophysical systems do not produce the combination of ring-like radio morphology, non-thermal radio spectrum, and detection of an optical continuum but not of optical emission lines. Probably the greatest difficulty with the lensing interpretation is the lack of positive evidence in the optical data. In a spectrum with no emission lines it is not possible to test for the presence of one or more extragalactic redshift systems. The optical data taken alone would suggest that MG1131+0456 is a radio galaxy, and the radio morphology could simply represent extremely unusual structure in the galaxy. When considering the plausibility of lensing, one must also remember that new classes of radio sources have been discovered by the VLA[9-13], and MG1131+0456 may, of course, be yet another example of such a discovery. The geometrical properties (though not the radiative properties nor the angular scale) of the ring are similar to those of the luminous arcs reported recently[14,15], and some of the explanations, and their tests, proposed for the arcs may also pertain to the elliptical ring of MG1131+0456[16]. Measurements that may help to determine the nature of MG1131+0456 may be possible with present

technology: for example, high-resolution VLBI (very-long-baseline interferometry) imaging of the radio structure, a measurement of the redshift(s), a measurement of the polarization at other wavelengths, and scaled VLA mapping at different frequencies to reliably measure spectral indices. Perhaps such measurements will reveal the nature of this unusual radio source.

We acknowledge the expert assistance of the staffs of the National Radio Astronomy Observatory and the National Optical Astronomy Observatories, and the helpful comments of the referees. Some of the data processing was done by E. Fomalont under the National Science Foundation Supercomputer Access Program. This work was supported by grants from the National Science Foundation and the National Aeronautics and Space Administration. D.P.S. acknowledges support from an Exxon fellowship. The National Radio Astronomy Observatory is operated by Associated Universities, Inc., under contract with the National Science Foundation. The National Optical Astronomy Observatories are operated by the Association of Universities for Research in Astronomy, Inc., under contract with the National Science Foundation.

Received 22 January; accepted 12 April 1988.

1. Einstein, A. *Science* **84**, 506 (1936).
2. Bennett, C. L., Lawrence, C. R., Burke, B. F., Hewitt, J. N. & Mahoney, J. *Astrophys. J. Suppl. Sev.* **61**, 1-104 (1986).
3. IRAS Point Source Catalog (US Government Printing Office, Washington DC, 1985).
4. IRAS Small Extended Source Catalog (US Government Printing Office, Washington DC, 1985).
5. A. Niell, *thesis*, Cornell Univ. (1971).
6. Schramm, T. & Kayser R. *Astr. Astrophys.* **174**, 361 (1987).
7. Blandford, R. D. & Kochanek, C. S. *Astrophys. J.* **321**, 658-675 (1987).
8. Turner, E. L., Ostriker, J. P. & Gott, J. R. III *Astrophys. J.* **284**, 1-22 (1984).
9. Yusef-Zadeh, F., Morris, M. & Chance, D. *Nature* **310**, 557-561 (1984).
10. Morris, M. & Yusef-Zadeh, F. *Astr. J.* **90**, 2511-2513 (1985).
11. Shaver, P. A., Salter, C. J., Patnaik, A. R., van Gorkom, J. H. & Hunt, G. C. *Nature* **313**, 113-115 (1985).
12. Becker, R. H. & Helfand, D. J. *Nature* **313**, 115-118 (1985).
13. Yusef-Zadeh, F. & Bally, J. *Nature* **330**, 455-458 (1987).
14. Soucail, G., Fort, B., Mellier, Y. & Picat, J. P. *Astr. Astrophys.* **172**, L14-L16 (1987).
15. Lynds, R. & Petrosian, V. *Bull. Am. astr. Soc.* **18**, 1014 (1986).
16. Paczyński, B. *Nature* **325**, 572 (1987).

A laser transition based on fluctuations

J. H. Hannay* & D. Field†

* H. H. Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK
† School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK

An unexpected phenomenon emerges from the model equations governing single-mode homogeneously broadened lasers. We have identified conditions analytically under which the laser output can be made to switch from low to high (or high to low) intensity through a small variation in the parameters of the medium, for example, the pump rate. Our results are distinct from those involving transitions at a 'second laser threshold' to regions of bistability, multistability and chaos[1,2]. The transition described here arises from the inclusion in our analysis of fluctuating forces in the medium[3]. A similar scheme can be used to describe a familiar transition in physics—the Curie–Weiss ferromagnetic transition.

By using the usual 'slowly varying envelope' and 'rotating-wave' approximations[4] the dynamics of a quantum two-level atom in a resonant classical electric field are described by the Maxwell-Bloch equations:

$$\dot{P} = -\gamma_p P + ED + f(t) \qquad (1a)$$

$$\dot{D} = -\gamma_D D + \lambda - EP \qquad (1b)$$

$$\dot{E} = -\gamma_E E + P \qquad (1c)$$

where P, D and E are dimensionless variables representing, respectively, the polarization, population inversion of the atoms and the electric field. They incorporate the physical constants h and ε_0, the atomic dipole moment and the resonant frequency. The γs are the associated decay constants (γ_p^{-1} and γ_D^{-1} corresponding to T_2 and T_1 in nuclear magnetic resonance), and λ is the inversion pump rate. Without the extra random noise term $f(t)$ discussed below, the equations constitute the celebrated 'Lorentz equations' of dynamical systems theory[5]. The steady-state solution exhibits the standard lasing transition: $E=0$ if $\lambda < \gamma_P\gamma_D\gamma_E$ and $E=(\lambda/\gamma_E - \gamma_P\gamma_D)^{1/2}$ if $\lambda > \gamma_P\gamma_D\gamma_E$. The inclusion of noise though is an important addition; the E values just given, for instance represent constant and therefore monochromatic fields with zero linewidth, an unphysical feature rectified by noise. It turns out that noise also has a different and rather surprising effect which is the subject of this work.

To obtain analytic results simplifications are necessary and we shall assume the decay and noise terms in the $\dot{P}$ equation dominate analogous terms in the others. Noise terms in the other equations have already been omitted, and we shall take the limit γ_D, γ_E, $\lambda \to 0$, with fixed ratios λ/γ_D, γ_E/γ_D. The noise will be assumed to be gaussian white noise with $\langle f(t)f(t+\tau)\rangle = \alpha\delta(\tau)$, where α measures its strength. We shall derive the statistically steady-state of equation (1): that is, the time-independent probability density $\rho(E)$, and in particular the average intensity $\langle E^2\rangle$. Our analysis will be exact, throughout, subject to these assumptions.

The first step is to identify that particular combination of E and D, namely $C = D + \frac{1}{2}E^2$, whose rate of change tends to zero in the stated limit, unlike those of E and D individually. From (1b) and (1c) we obtain

$$\dot{C} = -\gamma_D[C - (\tfrac{1}{2} - \gamma_E/\gamma_D)E^2 - \lambda/\gamma_D] \qquad (2)$$

whose solution is an integral from the present time t back over all past times $t-\tau(\tau>0)$

$$C(t) = \gamma_D\int_0^\infty e^{-\gamma_D\tau}[\lambda/\gamma_D + (\tfrac{1}{2} - \gamma_E/\gamma_D)E(t-\tau)^2]\,d\tau \qquad (3)$$

In the limit $\gamma_D \to 0$ this formula is recognized as a mathematical definition of the average, over all past time, of the square bracket. Thus, as expected, $C(t)$ does not vary in time; it is a constant whose value is determined by $\langle E^2\rangle$.

$$C = \lambda/\gamma_D + (\tfrac{1}{2} - \gamma_E/\gamma_D)\langle E^2\rangle \qquad (4)$$

We will now show that $\langle E^2\rangle$ is itself determined by C. Substituting $D = C - \frac{1}{2}E^2$ and $\dot{E} = P$ (from equation (1c) with $\gamma_E \to 0$) into the first Maxwell-Bloch equation we obtain the Langevin-type equation[5]

$$\ddot{E} + \gamma_P\dot{E} - E(C - \tfrac{1}{2}E^2) = f(t) \qquad (5)$$

This represents a fluctuating, damped motion of E in a cubic force field or quartic potential which has a double well shape for $C>0$. The joint probability distribution $\mathcal{P}(\dot{E}, E, t)$ that it generates is governed by the associated diffusion or 'Fokker-Planck' equation[5]

$$\partial\mathcal{P}/\partial t + \partial/\partial\dot{E}[(E(C - \tfrac{1}{2}E^2) - \gamma_p\dot{E})\mathcal{P}] + \partial/\partial E(\dot{E}\mathcal{P}) - \tfrac{1}{2}\alpha\partial^2\mathcal{P}/\partial\dot{E}^2 = 0 \qquad (6)$$

The steady-state solution is $\mathcal{P} = (\gamma_p/\pi\alpha)^{1/2}\exp(-\gamma_P\dot{E}^2/\alpha)\rho(E)$ where $\rho(E)$ is the probability distribution of E, a Boltzmann-like distribution in the quartic potential:

$$\rho(E) = N\exp[-\sqrt{(2\gamma_P/\alpha)}(\tfrac{1}{8}E^4 - \tfrac{1}{2}CE^2)] \qquad (7)$$

where N is the normalization constant. Thus the probability distribution of E itself depends through C on its own mean square value $\langle E^2\rangle$. Evaluating now the mean square value in terms of C we have

$$\langle E^2\rangle = \int E^2\rho(E)\,dE = \sqrt{(\cdot)}^{-1}D_{-3/2}(\sqrt{(\cdot)}C)/D_{-1/2}(\sqrt{(\cdot)}C) \qquad (8)$$

where $\sqrt{(\cdot)}$ stands for $\sqrt{(2\gamma_P/\alpha)}$ and the Ds are parabolic cylinder functions[6].

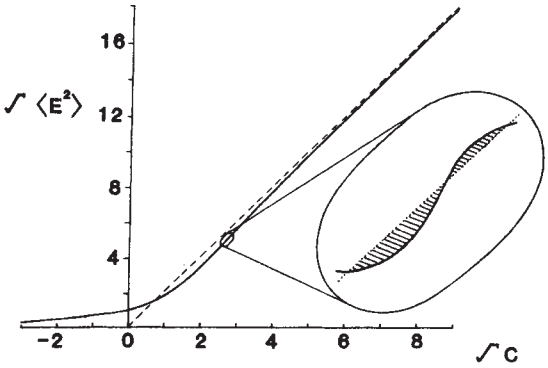


Fig. 1 The function $D_{-3/2}(-\sqrt{(\cdot)}C)/D_{-1/2}(-\sqrt{(\cdot)}C)=\sqrt{(\cdot)}\langle E^2\rangle$ against $\sqrt{(\cdot)}C$. The dashed line is the asymptote $2\sqrt{(\cdot)}C$. Analytic solutions of the Maxwell–Bloch equations are given by the intersections of this function with a straight line as described in the text. The inset sketches the inflection region and shows that there can be three solutions. The region with the larger neighbouring shaded area is the globally stable solution (the inner one here).

Solving equations (4) and (8) simultaneously we obtain self-consistent solution(s) for the average intensity $\langle E^2\rangle$. They can be displayed graphically as the intersections of the curve (8) and the straight line (4). In Fig. 1, $\sqrt{(\cdot)}\langle E^2\rangle$ is plotted against $\sqrt{(\cdot)}C$ so that the curve is simply the ratio of the parabolic cylinder functions (with its asymptote $\sqrt{(\cdot)}\langle E^2\rangle=2\sqrt{(\cdot)}C$ shown as the dashed line) and the straight line is $\sqrt{(\cdot)}\langle E^2\rangle=\sqrt{(\cdot)}C+B$ where $A=(\frac{1}{2}-\gamma_E/\gamma_D)^{-1}$ and $B=-\sqrt{(\cdot)}A\lambda/\gamma_D$ (shown only in inset).

Clearly there are positions of the straight line where there are zero, one or two solutions, although they may not be stable. Any intersection at which the gradient of the line lies between the gradient of the curve and zero is an unstable solution and will not be adopted by the differential equations (1*a*–*c*). So in a pair of solutions one is always unstable. The curve, however, has an inflection point (see exaggerated inset to Fig. 1) and therefore there is a narrow range of positions of the straight

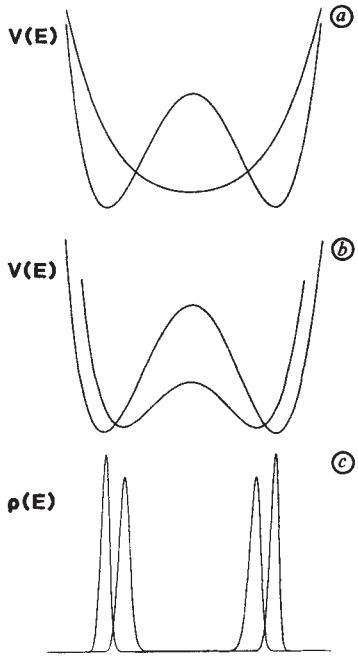


Fig. 2 *a*, The competing quartic 'potential' wells, $V(E)$, associated with lasing (double well) and non-lasing (single well). *b*, The competing quartics associated with the transition discussed here are both double wells but of different shapes. *c*, The two probability distributions $\rho(E)$ for the field E appropriate to the two double wells of (ii). One has larger mean intensity $\langle E^2\rangle$ than the other.

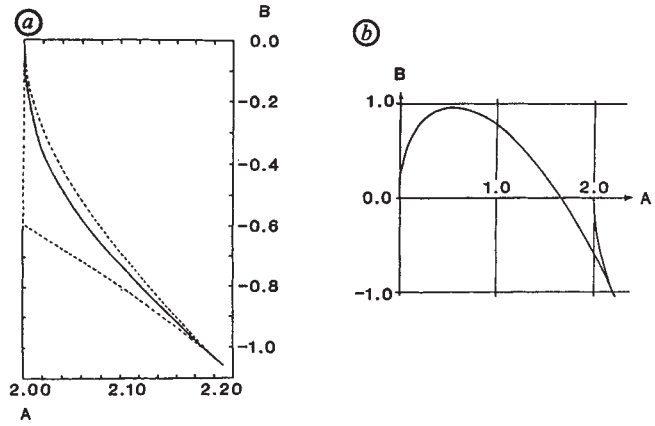


Fig. 3 *a*, The locus (solid line) of parameters $A=(\frac{1}{2}-\gamma_E/\gamma_D)^{-1}$ and $B=-\sqrt{(\cdot)}A\lambda/\gamma_D$ across which the transition in laser intensity occurs. The locus is defined by the gradient A and intercept B of the straight ine in the inset to Fig. 1 when drawn so as to give equal lobe areas. The cusp-shaped region of A,B space in which there are three intersections of the straight line with the curve is shown dashed. Outside this region there are fewer intersections and transition is not possible *b*, The full A,B space showing the continuation of the dashed lines in Fig. 3*a*. The cusp shaped region of triple intersection lies at the lower right-hand side of the diagram.

line with three intersections. Of these, the outer two are locally stable and the middle one is unstable. Local stability, though, is not enough because the Langevin equation will eventually seek out and adopt whichever of the two possibilities is globally stable. This follows if we invoke the fact that Langevin equations mathematically simulate thermal equilibrium[5]. The noise term simulates the action of a heat bath at an artificial temperture $(=1/k\sqrt{(\cdot)}$ here as can be seen from the Boltzmann-type distribution (7)). Thermal equilibrium adopts globally stable rather than locally stable or 'metastable' states and the simple rule distinguishing them is the Maxwell area rule, which we shall therefore apply here. It states that the intersection with the larger neighbouring lobe area (see inset to Fig. 1) is globally stable and the other is metastable. The implication of this rule is that by a small or, in principle, infinitessimal change in the parameters A or B, such that the change switches which lobe area is the larger, one can induce a large change in the average intensity $\langle E^2\rangle$ of the laser from one solution to the other or the reverse, there being, in principle, no hysteresis. Each of the two average intensities has an associated double-well quartic potential and the transition is, therefore, between two different symmetrically shaped double wells (Fig. 2*b*). This may be contrasted with the standard lasing transition which is between a single and a double well (Fig. 2*a*). In thermodynamic terminology our transition is first-order rather than the second-order lasing transition. Figure 2*c* shows the forms of the probability distributions (7) corresponding to the two potentials of Fig. 2*b*.

Figure 3*a* is the analogue of a thermodynamic phase diagram showing the locus of points A, B for which the straight line cuts equal lobe areas, and across which, therefore, this transition occurs. The locus lies, of course, within the small cusped region of the entire A, B space (Fig. 3*b*) where there are three intersections and terminates in a 'critical' cusp point with coordinates (2.21, −1.06) which are the A and B of the straight line tangent to the curve at its inflection point (2.76, 5.05) in Fig. 1. The other end (2.0, 0.0) of the transition locus corresponds to the straight line superimposed on the asymptote line and as such is an especially singular point, not a standard cusp point. For the entire region $A>2$ in which the transition locus lies there is one globally stable solution. The region $0<A<2$ is not physically accessible because it requires $\gamma_E/\gamma_D<0$. The region $A<0$ has again one globally stable solution corresponding to the single intersection point.

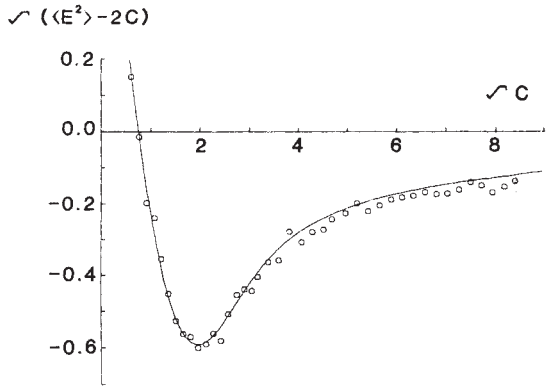


Fig. 4 The deviation of $\sqrt{(\cdot)}\langle E^2\rangle$ from the asymptotye line $2\sqrt{(\cdot)}C$ show in Fig. 1 (solid line). The open circles are results of numerical solutions of the Langevin equation (5) itself based on the premise that λ, γ_D, $\gamma_E \to 0$. Numerical solutions were obtained with $A = 2.0$ (corresponding to $\gamma_E/\gamma_D = 0$), $B = -0.3125$. (Values of $A > 2.0$ corresponding to $\gamma_E/\gamma_D \neq 0$ yield equally good agreement with the predictions of (8).) These results therefore demonstrate that the analysis leading from (5) to (8) has been correctly done and shows, as (8) predicts, that the curve shown is universal, the values of γ_P and the noise strength being immaterial.

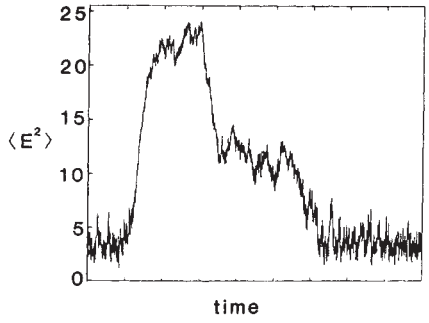


Fig.5 The temporal response of $\langle E^2\rangle$ calculated from equation (1) with $A = (\frac{1}{2} - \gamma_E/\gamma_D)^{-1} = 2.05$ (corresponding to $\gamma_E/\gamma_D = 0.0122$), $B = -\sqrt{(\cdot)}A\lambda/\gamma_D = -0.446$ initially, changing abruptly to -0.646 and then back again. These calculations show the existence of the laser transition region as predicted by equation (8).

To test our analytical theory we have performed numerical simulations, with the noise supplied by a random number generator. Firstly, the constancy of C was checked. Then the curve (8) relating $\sqrt{(\cdot)}\langle E^2\rangle$ and $\sqrt{(\cdot)}C$ was verified by solving (5) numerically. To display this clearly, it is convenient to plot the deviation of $\sqrt{(\cdot)}\langle E^2\rangle$ from the asymptote line, that is $\sqrt{(\cdot)}(\langle E^2\rangle - 2C)$ against $\sqrt{(\cdot)}C$. In Fig. 4 the solid line is the analytical deviation and the open circles are the numerical verification. To show the time response of the intensity $\langle E^2\rangle$ to a change in the parameters, the full set of equations (1) was then solved with $\gamma_D = \gamma_P/10$. Keeping A constant, the value of B was changed by changing the pump rate λ so that the straight line in the graphical solution moved to switch the intensity from low to high and then back again. Figure 5 shows the time response of the intensity calculated numerically.

It is not yet clear how robust the transition effect is to elaboration of the model that has been used. There is room for improvement in the treatment of the phase fluctuations whose effect was crudely modelled by the white noise in the polarization equation (1*a*). Indeed it may no longer be justified to treat the variables P, D, E as real in the presence of these fluctuations. Also the noise terms associated with γ_D and γ_E were taken as zero. If these terms are not zero, even if they are only small, their influence should strictly be included. One common simplification of the model, however, the 'rate' approximation, in which $\dot{P}$ is set equal to zero, does not affect our result in any way. It simply erases the $\ddot{E}$ in (5) and, though the consequent Fokker-Planck equation is changed, the steady-state solution (7) is not.

Finally the application of the method used above to a more familiar (but less interesting) problem in physics may be worth mentioning. It is the Curie-Weiss model of ferromagnetism[7]. The magnetic moment m, (one component of a classical dipole moment vector) can be considered as subject to a steady 'force' $-c$ (proportional to the magnetic field) plus a thermal noise term so that $\dot{m} = -c + f(t)$ analogous to (1*a*). This Langevin equation yields a Fokker-Planck equation whose steady-state solution is, as expected, a Boltzmann exponential distribution $N\exp(-cm\sqrt{2/\alpha})$. But the constant c is itself determined by $\langle m\rangle$. In fact, $c = A\langle m\rangle + B$, where B corresponds to an external magnetic field and $A\langle m\rangle$ to the internal one generated by the magnetization itself. This relation can be induced analogously to (2) by liberating c and giving it artificial dynamics $\dot{c} = -\gamma(c - Am - B)$, where $\gamma \to 0$. Numerical solution of these differential equations for m and c confirms the transition expected from the Maxwell area rule: as B is changed through zero the value of $\langle m\rangle$ suddenly flips to keep the same sign as B. In summary, both the Curie-Weiss system and the Maxwell-Bloch system share the property that they yield a probability distribution which depends on its own mean or mean square. It is this unfamiliar self-consistency which in both cases leads to a discontinuous response to a smoothly varying influence: in one the external magnetic field, in the other (say) the pump rate.

Received 7 August 1987; accepted 5 April 1988.

1. Abraham, N. B. Lugiato, L. A. & Narducci, L. M. *J. opt. Soc. Am. B*2, 7-14 (1985).
2. Haken, H. *Light vol. 2, Laser Light Dynamics* (North Holland, Amsterdam, 1985).
3. Pike, E. R. & Sarkar, S. (eds) *Frontiers in Quantum Optics* (Hilger, New York,, 1986).
4. Sargent, M., Scully, M. & Lamb, W. *Laser Physics* (Addison Wesley, New York, 1974).
5. Haken, H. *Synergetics* (Springer, Berlin, 1983).
6. Abramowitz, M. & Stegun, I. A. *Handbook of Mathematical Functions* (Dover, New York, 1970).
7. Tabor, D. *Gases, Liquids and Solids* (Cambridge University Press, 1969).

Imaging of liquid crystals using a tunnelling microscope

J. S. Foster & J. E. Frommer

IBM Research Division, Almaden Research Center, San Jose, California 95120, USA

Since its conception[1], the scanning tunnelling microscope (STM) has been used principally to explore surfaces of well ordered inorganic solids and graphite[2,3]. We now report the first use of the STM to unambiguously image well defined organic adsorbates on surfaces. We have imaged individual organic molecules in a liquid crystal array on a graphite surface with the STM with nearly atomic resolution. Liquid crystals have previously been analysed by X-ray diffraction, neutron scattering, dilatometry and other means to determine molecular order. Here we identify a two-dimensional order in 4-*n*-octyl-4′-cyanobiphenyl, a smectic in the bulk, confirming and embellishing literature models. We have also imaged 4-(*trans*-4*n*-pentylcyclohexyl)benzonitrile, a nematic liquid crystal in the bulk at room temperature. We observe an increase in order in both liquid crystals with respect to their bulk phases. The bulk smectic appears crystalline on the graphite surface, as the smectic planes are perfectly registered with one another. The structure of the bulk nematic at the interface with the surface shows a strong conformity to the graphite substrate. In both cases the molecular axes lie parallel to the substrate, and the STM images a cross-section along the direction of the molecular axes of the classical bulk phases (Fig. 1).

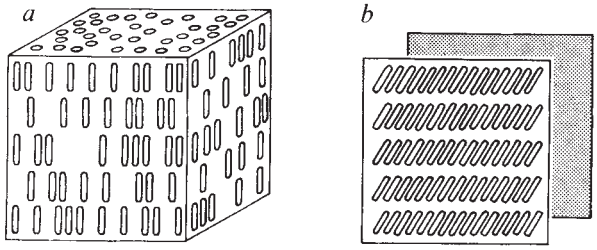


Fig. 1 *a*, Generalized rendition of smectic ordering in liquid crystals showing multiple planes in cross-section. Note that there is no registration between molecules of adjacent planes. *b*, Proposed model showing liquid crystal molecules on a graphite surface (fine ordered array of dots), molecular axes lying parallel to the surface and registered between planes.

In scanning tunnelling microscopy[4], a conducting needle is held in close proximity to a conducting surface. As the needle approaches a spacing ~10 Å from the surface, electrons can tunnel with a reasonable probability between the needle and surface. A voltage bias is applied between the two conductors and the resulting current is a sensitive measure of their separation. In this study, a tungsten tunnelling needle is laterally scanned across the surface and the current is held constant by driving the tip perpendicularly to the sample plane, using an applied voltage on a piezoelectric element. This piezoelectric voltage is measured during scanning and translated into an image. For example, a relatively low conductivity (or topographic) region would require the tip to translate towards the surface to keep the current constant; this would appear as a dark region in imaging. Conversely, light areas correspond to higher relative conductivities (or topographies) between the tip and substrate.

Our samples are prepared by applying a drop of the liquid crystal to a freshly cleaved single crystal of graphite (HOPG, highly oriented pyrolitic graphite). The tip is immersed in the drop and mechanically guided towards the conducting graphite surface. At this point, a low-voltage bias (30 mV) can be used to image the graphite with atomic resolution. The graphite appears as a perfectly regular lattice and no defects are apparent[2,3]. To image these liquid crystals, it is first necessary to wait several hours after deposition of the liquid. This requirement may stem from the slow process of relatively large smectic arrays self-ordering on the graphite surface. Experiments are in progress to control sample temperature; relatively rapid ordering should be observed from a sample drop cooled from the isotropic liquid phase. Imaging of the liquid crystal is performed with a direct current bias from the tip to the sample of −0.5 to −0.7 V and a tunnelling current of 0.2 to 0.5 nA. We have no quantitative measure of the thickness of the imaged liquid crystal films; from the magnitude of the tunnelling current, we estimate it to be between one and several monolayers.

Figure 2 shows the liquid crystal 4-*n*-octyl-4′-cyanobiphenyl (8CB) imaged by the STM at room temperature. These images are stable for days during imaging (tunnelling continuously). The scale of Fig. 2*a*, 215 × 215 Å, allows the visualization of the interplanar packing and demonstrates long-range order. The periodicity of this interplanar packing is ~34 Å. Figure 2*b* focuses on a smaller scale, 75 × 75 Å, and allows the resolution of individual molecular axes spaced ~5 Å apart. These axes form an angle of ~35° with respect to the interplanar packing. The molecules in one plane register with the molecules in adjacent planes. A third periodicity appears within the planes as repetitive hole-like features at a spacing of every four molecules (that is, 20 Å). We also note that there is a registry of these hole-like features from plane to plane, consistent with the interplanar registry between molecular axes. During these experiments, several structures were seen (manuscript in preparation).

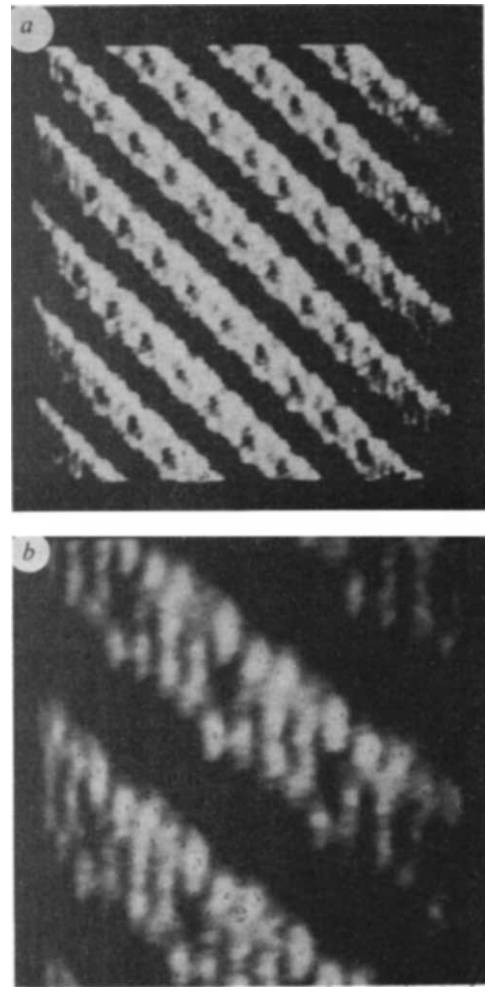


Fig. 2 *a*, STM image of 8CB on a graphite substrate. The scale is 215 × 215 Å. The image shows the two-dimensional structure of the liquid crystal. *b*, Higher magnification view of *a*, 75 × 75 Å.

These different structures could be a manifestation of exact thermal histories or deposition conditions.

Whereas 8CB exists in the smectic A phase in the bulk at room temperature[5-8], a second liquid crystal, 4-(*trans*-4-*n*-pentylcyclohexyl)benzonitrile (5CBN), exists in the supercooled nematic phase in the bulk under the ambient conditions in which the STM data are recorded[9]. STM images of 5CBN are shown in Fig. 3. Figure 3*a*, 165 × 165 Å, displays an overall one-dimensional order in which all molecular axes run in a common direction. Figure 3*b* shows a higher magnification (27 × 27 Å) of a region displaying two-dimensional order. Three features of Fig. 3*b* are particularly notable: (1) a spacing of 2.5 Å between the individual bright spots that constitute the lines—highly suggestive that the spots correspond to individual (and alternating) atoms along the molecular axes; (2) a spacing of 4.2 Å between the lines—the apparent separation between the molecules; and (3) a registration between atoms in adjacent chains that forms an axis perpendicular to the molecular axis. These three features are consistent with a picture of the molecules superimposed on and registered with the graphite lattice. The 2.5 Å dimension correlates with the spacing between alternating carbons in an alkyl chain, as well as with the spacing between every other carbon in the graphite lattice. (The spacing between alternating carbons of graphite and the hydrocarbon fortuitously coincide within 2%, because the shorter graphite bond is compensated for by a larger bonding angle.) The 4.2 Å dimension observed between molecules corresponds to the distance between alternating hollow sites in the graphite lattice,

and is reasonable as a packing distance between the adsorbed molecules. Finally, the right angle formed between the molecular axes and the intermolecular axis is consistent with the graphite geometry.

Figure 3*c* shows a three-dimensional view of the nematic liquid crystal, 72×72 Å. The discontinuity in the centre of Fig. 3*c* represents a boundary between two different orientations of the liquid crystal. These two orientations are not exclusive: several others have been observed, and there is a correspondence with the natural lattice vectors of the surface of graphite. The angle between these two domains is 30°, implying that the adsorbed organic is sensing the sixfold symmetry of the graphite substrate. This is particularly interesting as the STM normally images only threefold symmetry on graphite. The discontinuity noted in the centre of Fig. 3*c* extends in the zig-zag pattern for hundreds of ångströms; its significance is unknown.

In addressing the packing arrangement of the liquid crystal molecules, we start by considering the structures that have already been put forward[5-8]. The longest periodicity that we observe in 8CB is the interplanar spacing of about 34 Å, compared to 29-32 Å observed by X-ray scattering. Surface forces from graphite could account for this discrepancy; changes in periodicity have been previously observed on application of external forces such as increasing pressure[6] or decreasing temperature[10]. One feature which is consistent in the literature is that 8CB is believed to exist in an interdigitated smectic-A bilayer because the detected 30 Å periodicity is greater than the length of the molecule (22 Å in a linear conformation), yet less than twice the length of the molecule. From the STM images, the 34 Å periodicity comprises a light 20 Å-wide 'layer' of close-packed parallel molecules and a dark interlayer 'space' 14 Å wide. A reasonable explanation is that the dark regions are occupied by the saturated aliphatic half of the molecule, whereas the light regions are occupied by the conjugated cyanobiphenyl groups. It is reasonable to expect that the tunnelling current will be different for these two sections. For example, if the tunnelling current is enhanced through the conjugated head relative to the saturated tail, then the heads would appear brighter in the images. This effect would be enhanced on tunnelling through additional layers of these molecules, depending on the specific orientation between the layers. This explanation invokes a correlation between electrical conductivity and the probability of electron tunnelling. The electrical conductivity of the conjugated moiety is expected to be enhanced due to the delocalized π-electron density of the phenyl and cyano groups; this effect has been widely studied in the field of electronically conducting polymers whose conductivity is predicated on extended π-electron conjugation[11,12].

An alternative explanation for the long range periodicity recorded by the STM is that the 20 Å dimension represents the length of a singular molecular axis that is bent at the juncture between the aromatic and aliphatic portions. The dark interlayer 'space', a region of decreased tunnelling current, could be a topographical effect resulting from the way the molecules pack in three dimensions. In this context, the faint features seen in the dark regions would originate from a molecular layer beneath the one being imaged in the bright regions. A similar topographical effect could account for the hole-like features of Fig. 2*b*. They could arise from molecular ordering effects from surface-adsorbate interactive forces (such as incommensurate packing) or from intermolecular forces between the polar and non polar ends of adjacent molecules along a series of aligned molecules. The latter could lead to an increasingly repulsive force, resulting in a periodic dislocation of the series to restore the array to a lower-energy conformation. Finally, the hole-like features could originate from periodic inclusion of solvent molecules, the periodicity remaining even after evaporation of the solvent[13].

We have reported the imaging of liquid crystals using scanning tunnelling microscopy. Periodicities of the order of 5 Å and 34 Å are observed for 8CB, corresponding to known ordering of the molecules within the planes and between the planes respectively. An additional periodicity of 20 Å has been observed within the planes; this third periodicity has not been previously reported. We believe that we can visualize the orientation of the molecular axes relative to the two-dimensional ordering in the liquid crystal. In the images of the 5CBN, the interatomic and intermolecular distances and angles strongly suggest that the graphite substrate directs orientation of the adsorbed molecules. Finally, although our exploratory studies with the STM are only just beginning to reveal molecular arrangements, we believe this technique has enormous potential.

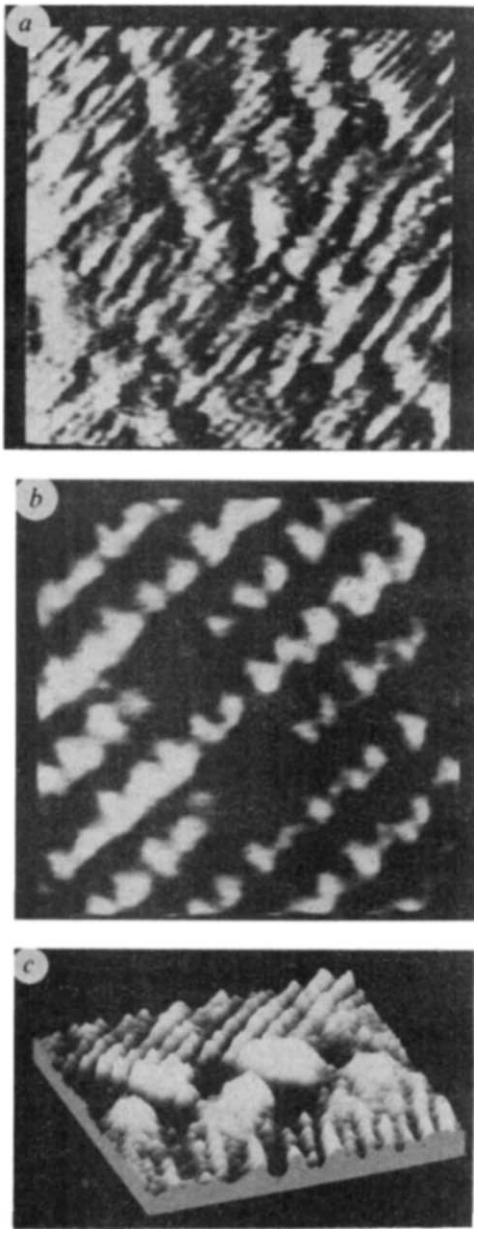


Fig. 3 *a*, STM image of 5CBN on a graphite substrate. The scale is 160×160 Å. *b*, Higher magnification view, 27×27 Å. The spacing between bright spots in the chains is about 2.5 Å, and about 4.2 Å between chains. *c*, A three-dimensional view of the boundary between two orientations (30° rotation) of 5CBN. The scale is 72×72 Å. The figure is obtained by assigning a vertical height from the data, as well as a brightness on each line scan in a rastor.

We thank R. Twieg, E. Barrall, M. Dovek and C. Lang for their assistance and L. J. LaComb for help with the image processing.

Received 22 February; accepted 29 April 1988.

1. Binnig, G. & Rohrer, H. *Helv. phys. Acta* **55**, 726-735 (1982).
2. Park, S. I. & Quate, C. F. *Appl. phys. Lett.* **48**, 112-114 (1986).
3. Batra, I. P. *et al. Surf. Sci.* **181**, 126-138 (1986).
4. Hansma, P. K. & Tersoff, J. *J. appl. Phys.* **61**, R1-R24 (1987).
5. Lydon, J. E. & Coakley, C. J. *J. phys. Colloq.* **36**, C1-45 (1975).
6. Cladis, P. E., Bogardus, R. K., Daniels, W. B. & Taylor, G. N. *Phys. Rev. Lett.* **39**, 720-723 (1977).
7. Leadbetter, A. J., Durrant, J. L. A. & Rugman, M. *Molec. Cryst. liq. Cryst.* **34**, 231-235 (1977).
8. Leadbetter, A. J., Frost, J. C. & Gaughan, J. P. *J. Physique* **40**, 375-380 (1979).
9. Pohl, L., Eidenschink, R., Krause, G. & Erdmann, D. *Phys. Lett.* **60A**, 421-423 (1977).
10. Ribotta, R. *J. Phys. Colloq.* **37**, C3-149 (1976).
11. Frommer, J. E. & Chance, R. R. *Encyclopedia of Polymer Science and Engineering* 2nd edn Vol. 5 462-507 (1986).
12. Kaner, R. B. & MacDiarmid, A. G. *Scient. Am.* **258**, 106-111 (1988).
13. Barrall, E. M. & Johnson, J. F. *Liquid Crystals and Plastic Crystals* Vol. 2 (eds Gray, G. W. & Winsor, P. A.) 254-306 (Wiley, New York, 1974).

The weather attractor over very short timescales

A. A. Tsonis & J. B. Elsner

Department of Geosciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201, USA

Recent work has used ideas from the theory of dynamical systems in the study of climate and weather over timescales ranging from decades to hundreds of thousands of years[1–5]. In this study, similar ideas are applied to weather observations over a time interval of 11 hours. The results suggest the existence of a low-dimensional strange attractor.

According to the theory of dynamical systems[6], the evolution of a system can be described by trajectories in the state space. The coordinates of the state space are defined by the variables needed to completely describe the evolution of the system. Each trajectory in the state space represents the evolution of the system from some initial condition. If the system exhibits an attractor, all trajectories initiated from different initial conditions will eventually converge and stay on a submanifold of the total available space. This submanifold 'attracts' the trajectories and it is called an attractor. For systems that develop deterministically, their attractors are low-dimensional smooth topological manifolds such as points, limit cycles and toruses. These attractors are, therefore, characterized by an integer dimension that is equal to the topological dimension of the submanifold in the state space. An important property of these attractors is that trajectories converging on them do not diverge, and thus stay at a constant distance from each other. This property guarantees long-term predictability of this system. But the trajectories do not have to stay on a smooth topological manifold. For many dynamical systems it has been found that the trajectories stay on an attracting submanifold that is not topological. These submanifolds are called 'fractal' sets and are characterized by a dimension that is not an integer[7]. The corresponding attractors are called 'strange' attractors. An important property of these attractors is the divergence of initially nearby trajectories. Such an action imposes limits on prediction[8]. Thus, long-term predictability for these systems is not guaranteed. Such systems are then deterministic only in the sense of being described by well behaved differential equations, but they are not periodic or quasi-periodic (even if they are not disturbed irregularly)[9]. Note that the dimensionality of an attractor, whether fractal or not, indicates the minimum number of variables present in the evolution of a system. Therefore, the determination of the dimension of an attractor sets a number of constraints that should be satisfied by a model used to predict the evolution of a system.

In the case where the exact mathematical formulation of the system is not available, the state space can be replaced by the phase space which can be produced using a single record of some observable variable from that system[10,11]. If this system is the atmosphere, for instance, then the observable variable could be the temperature or the pressure or the geopotential. Nicolis and Nicolis[1] were the first to apply the above ideas in climatic studies. Using single-variable values of the oxygen isotope records of deep-sea cores spanning the past million years, they reported a dimension for the climatic attractor equal to 3.1. Fraedrich[2,3] and Essex *et al.*[4] have also analysed weather data over timescales ranging from 15 to 40 years and reported a dimension of the weather attractor between 6 and 7.

Given an observable, say $x(t)$, if the attractor has finite

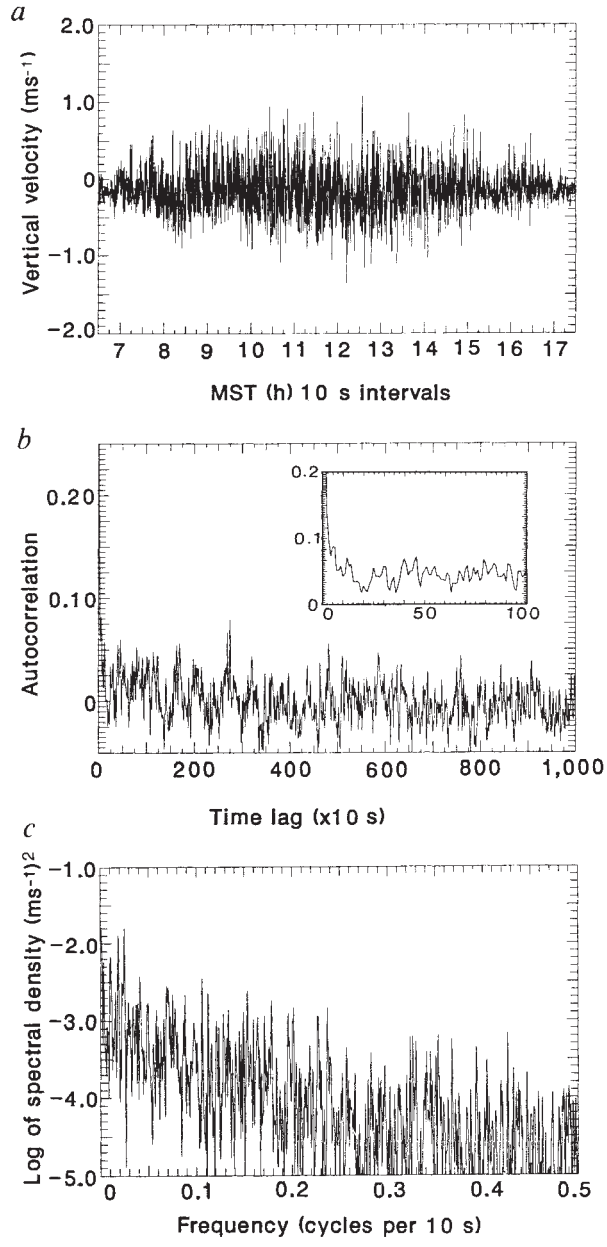


Fig. 1 *a*, The data used in this study represent 10-second averages of the vertical wind velocity over 11 hours. The data were recorded from 0630 to 1730 MST (Mountain Standard Time) on 26 September 1986 at Boulder, Colorado. At about 0630 the sun rises. The air close to the ground is heated and rises, creating strong convection. Positive values indicate updrafts and negative values indicate downdrafts. *b*, The autocorrelation function for the above data. The inset graph is a magnification of the region close to the origin. *c*, The logarithm of the spectral density as a function of the frequency for the above data. The spectra show various peaks on a background of a continuous frequency spectrum. This suggests that a strange attractor may be present.

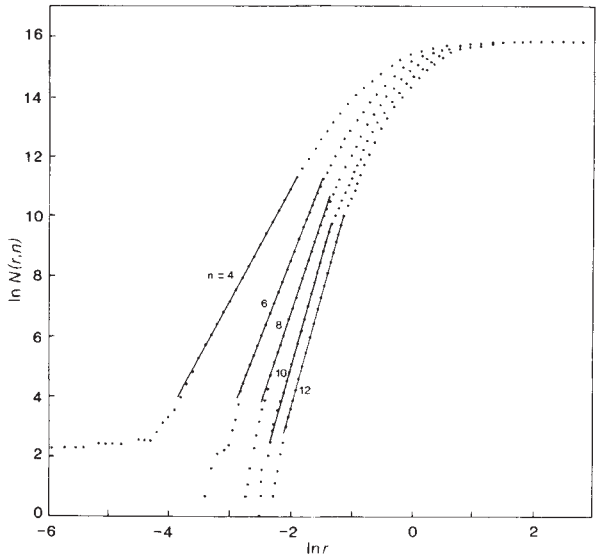


Fig. 2 Plot of ln $N(r, n)$ against ln r for embedding dimensions, $n = 4, 6, 8, 10, 12$. Note the convergence of slopes as n increases.

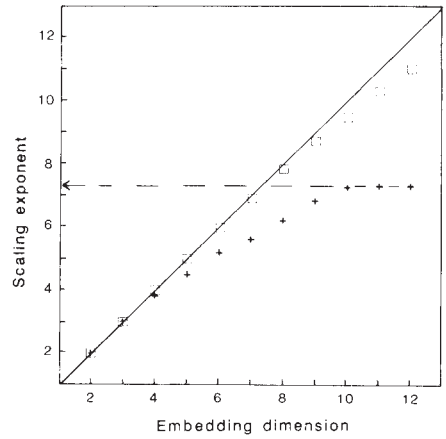


Fig. 3 Scaling exponent, d, as a function of the embedding dimension, n. Crosses correspond to the wind velocity data and squares to a random sample of the same size as the wind data. Note the saturation of the scaling exponent observed for the wind data, although there is no saturation for the random set. From this figure we estimated $d_s = 7.3$.

dimension the complete state vector $\boldsymbol{X}(t)$ can be constructed by the following process[10,11]: $x(t+\tau)$ is used as the first coordinate, $x(t+2\tau)$ as the second ... and $x(t+n\tau)$ as the last, τ being the delay parameter. Thus, for an n-dimensional phase space, a 'cloud' of points will be generated. From this 'cloud' the number of pairs can be found, $N(r, n)$, with distances less than a distance r. If for significantly small r we find that

$$N(r, n) \propto r^d \tag{1}$$

we call d the (correlation) dimension of the attractor for that n (ref. 12). We can then test equation (1) for increasing values of n (the embedding dimension of the phase space). If the value of d becomes independent of n (that is, it reaches a saturation value, d_s), this is evidence that the system represented by the time series should possess an attractor with a dimension equal to d_s; otherwise the time series is random.

It is important to note that when determining the dimension of the attractor (for any embedding dimension of the phase space), only pairs that are separated by a time interval greater than the decorrelation time should be considered. In addition, one should be very cautious in using very few actual data points or highly smoothed data: wrong conclusions can be drawn if the above-mentioned points are not taken into consideration[4,5].

In this work, 10-second averages of the vertical wind velocity recorded 10 metres above the ground over an 11-hour period are analysed. Thus, the total number of points is 3,960. The data are over a timescale much smaller than the previous analyses, and were recorded from 0630 to 1730 MST (Mountain Standard Time) on 26 September 1986 by the National Oceanic and Atmospheric Administration in Boulder, Colorado. Figure 1 shows the data, their autocorrelation and spectral density function. The decorrelation time may be defined as the lag time at which the autocorrelation falls below a threshold value. This threshold value is not uniquely defined, and in general it depends on the problem in hand and the assumptions about the data set. In meteorology this threshold value is commonly defined as $1/e$, especially if the autocorrelation function is nearly exponential[13]. Figure 1*b* (inset) indicates that the autocorrelation function decays nearly exponentially. But in view of the potential problems when correlated pairs are included in the calculations[4,5], we decided to be more conservative. We thus used a decorrelation time that is more 'restrictive' and at the same time allowed us to work with a sufficiently large number of pairs. We adapted a decorrelation time equal to 20 seconds. For time lags greater than 20 seconds, the autocorrelation drops and remains below a value of ~0.10. Figure 2 shows the logarithm of the number of pairs plotted against the logarithm of r for $\tau = 10$ seconds for selected embedding dimensions. From this information for each embedding dimension, the scaling region is determined and its slope is calculated by fitting a straight line in that region. The scaling region is usually observed in the region $3.0 < \ln N(r, n) < 11.0$. The slope of the straight line gives the scaling exponent d in equation (1). An excellent discussion about the scaling behaviour of a particular data set and the possible problems associated with fitting a straight line to intervals of r for which r is too large or too small is given by Essex *et al.*[4]. Figure 3 shows this scaling exponent as a function of the embedding dimension, together with a plot representing our time series as a random sample of the same size as our data set. As can be seen in Fig. 3, as n increases, $d \to d_s = 7.3$, but note that no such saturation is observed for the random sample. Therefore, we can conclude that the system represented by the vertical wind velocity series possesses an attractor. The non-integer dimensionality of the attractor indicates that the attracting submanifold is a fractal set and that the attractor is 'strange'.

In that case, the value of d_s suggests that over very short timescales the weather might be represented as a dynamical system with at least eight degrees of freedom (eight differential equations). We thus share the conviction of Nicolis and Nicolis[14] and Essex *et al.*[4] that the atmosphere exhibits properties of a low-dimensional attractor, even though our work refers to a timescale smaller than theirs.

Our analysis was focused on daytime data. During the night the forcing of the Sun is minimum and the wind data do not fluctuate significantly above a mean value of about zero. Thus during a 24-hour interval, the system switches back and forth between dynamical regimes of low and intense activity. Thus, if night-time data are included in the calculation of the correlation dimension, a large number of data points will cluster around one point in the phase space. The result will be an underestimation of the correlation dimension. The above problems are avoided when daytime data are used. The question then arises whether or not the weather attractor over night is of the same dimension as over day, and how the internal dynamics unravel from the day–night cycle. Work in this area is in progress.

It is interesting to speculate on the results reported above. In the early sixties, Lorenz[15] demonstrated that a dynamical system consisting of the three differential equations that give the description of a horizontal fluid layer heated from below exhibits a strange attractor. The warmer fluid formed at the bottom is

lighter and it tends to rise, creating convection currents. This phenomenon takes place in the Earth's atmosphere, where the air close to the ground is heated and rises. This monumental work provided the first mathematically based explanation for the unpredictability of weather. The data analysed here are strongly connected with convection, and the results suggest that in reality the dynamical system that will produce such data should be described by at least eight differential equations, thus being much more complicated and more unpredictable.

Our results support the existence of a low-dimensional attractor, but should be interpreted with caution. As shown in Osborne *et al.*[16], under certain circumstances a stochastic signal may be simulated with a similar structure as some observed data which (stochastic signal) may have a small, finite value for the correlation dimension, even when no underlying attractor is present. Even though this does not necessarily imply that the existence of a small, finite dimension in some observed data is not due to a strange attractor, we think that it is worth mentioning. We hope that research in this area will be beneficial to the debate about the possible existence of low-dimensional attractors in weather and climate or in other systems.

We thank Professors Peter Grassberger and Christopher Essex for discussion, the National Oceanic and Atmospheric Administration and the National Center for Atmospheric Research (NCAR) for providing us with the data and allowing us to use their CRAY computer for our computations, and Ken Hansen for computing assistance.

Received 14 October 1987; accepted 14 April 1988.

1. Nicolis, C. & Nicolis, G. *Nature* **311**, 529-532 (1984).
2. Fraedrich, K. *J. atmos. Sci.* **432**, 419-432 (1986).
3. Fraedrich, K. *J. atmos. Sci.* **44**, 722-728 (1987).
4. Essex, C., Lookman, T. & Nerenberg, M. A. H. *Nature* **326**, 64-66 (1987).
5. Grassberger, P. *Nature* **323**, 609-612 (1986).
6. Holden, A. V. (ed.) *Chaos: an Introduction* (Manchester University Press, 1985).
7. Mandelbrot, B. B. *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).
8. Crutchfield, J. P., Farmer, J. D., Packard, N. H. & Shaw, R. S. *Scient. Am.* **255**, 46-57 (1986).
9. Grassberger, P. in *Chaos: an Introduction* (ed. Holden, A. V.) 291-311 (Manchester University Press, 1985).
10. Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. *Phys. Rev. Lett.* **45**, 712-716 (1980).
11. Takens, F. *Lect. Not. Math.* **898**, 366-381 (1981).
12. Grassberger, P. & Procaccia, I. *Physica* **9D**, 189-208 (1983); *Phys. Rev. Lett.* **50**, 346-349 (1983).
13. Zawadzki, I. *J. appl. Met.* **12**, 459-472 (1973).
14. Nicolis, C. & Nicolis, G. *Nature* **326**, 523 (1987).
15. Lorenz, E. N. *J. atmos. Sci.* **20**, 130-141 (1963).
16. Osborne, A. R., Kirwan, A. D. Jr, Provenzale, A. & Bergamasco, L. *Physica* **23D**, 75-83 (1986).

A guide to Phanerozoic cold polar climates from high-latitude ice-rafting in the Cretaceous

L. A. Frakes & J. E. Francis

Department of Geology and Geophysics, University of Adelaide, GPO Box 498 Adelaide, South Australia, 5001

The high-latitude extent of warm-climate indicators at certain times in Earth history has been considered as evidence that the globe was ice-free for long intervals, despite theoretical considerations and results from numerical modelling experiments[1,2] indicating that this was unlikely. One of the warmest periods, the Cretaceous, displays faunal and floral evidence for 'cool-temperate' to 'sub-tropical' conditions very near to the poles. However, our studies of Lower Cretaceous mudstones of central Australia that contain outsized exotic blocks have led to the conclusion that the blocks were emplaced by ice-rafting, implying that high-latitude ice was present at sea level. Strata of a similar origin of mid-Jurassic to mid-Cretaceous age occur on other continents that were positioned between 65° and 78° palaeolatitude. Indeed, there is a record of high-latitude ice-rafting throughout the Phanerozoic, suggesting that ice was present on Earth for much of its history, and that ice-free conditions could have been at most only episodic over the past 600 Myr.

During the early Cretaceous, central Australia was positioned in mid to high latitudes[3] and was occupied by a large intracratonic basin, the Eromanga Basin, in which marginal sandstones and marine mudstones (the Bulldog Shale) were deposited. The Bulldog Shale of the southwestern Eromanga Basin (Valanginian to Albian in age) is a dark bioturbated or laminated shale with up to 10% sand in its matrix and occasional sandy and silty laminae. It contains large, exotic clasts of mainly Precambrian quartzite and volcanic rock up to 3-m diameter. Clasts occur both as lonestones and as scatterings or concentrations along bedding planes. They show a high degree of rounding and were probably derived from river and/or shoreline environments.

The hydrodynamic paradox presented by the occurrence of outsized boulders in these mudrocks can only be resolved if the clasts were transported laterally into the basin by swift currents or mass-movement processes such as mud flows, or dropped at the depositional site from a floating raft. But strong currents competent to carry large clasts also scour the substrate and deposit layers of sand and coarser material as turbidites, which include erosional surfaces and widespread coarse layers exhibiting graded bedding. Mudflows produce massive non-laminated structures[4]. In contrast, rafted facies are ideally defined by bedding penetrated by the fallen clast; this may only be apparent on the millimetre scale and is often obscured by compaction of mud around the clast. Another important feature of ice-rafted detritus is the sand-sized component known to be present in modern icebergs[5].

Evidence of mass-movement in the form of turbidites or mudflows in the Eromanga Basin is lacking however. Moreover, penetration structures in the laminae below the clasts suggest that the boulders are dropstones, in that they fell vertically from rafts[6], rather than being transported laterally by currents. From sedimentary evidence within the Bulldog Shale, including the presence of glendonites[7-9] and quartz sand grains with surface textures characteristic of glacial regimes, we conclude that ice acted as the rafting mechanism that transported the boulders out into the basin. Although fossil wood is present in the shale and we have considered it as a rafting agent, the large size of

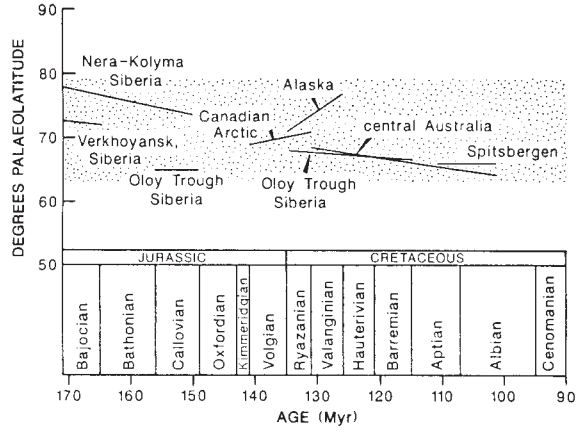


Fig. 1 Distribution of late Mesozoic ice-rafted deposits by palaeolatitude and age. Uncertainties in dating and palaeomagnetic measurements are included in the length of the line for each region. Stippled area (~65-78°) represents full distribution of reported ice-rafted deposits. Palaeolatitudes derived from ref. 26.

many of the clasts (up to 3 m in diameter), and the large volume of wood required to carry them, makes this an unlikely method of emplacement. Lonestones we interpret as rafted stones transported by marine shore ice or river ice; the layers of clasts we consider to be the result of reworking and winnowing of clast-bearing muds by storm waves and currents. Ice-rafting apparently occurred only in the southwestern part of the Eromanga Basin because clasts in the Bulldog Shale are only recorded from northern South Australia and southwestern New South Wales. This part of the basin was located at palaeolatitudes higher than 65°.

On a global scale, late Mesozoic sediments containing ice-rafted clasts have been described from several sites, all located in high palaeolatitudes and of mid-Jurassic to mid-Cretaceous age (Fig. 1). The most widespread occurrence of Mesozoic ice-rafted deposits is recorded from the USSR between Kamchatka and western Siberia (Oloy Trough, Verkhoyansk region and Nera-Kolyma interfluve)[10]. Other notable areas include the Sverdrup Basin in the Canadian Arctic[11] and Spitsbergen[12]. Lower Cretaceous shales with outsized clasts are also reported from the northern Brooks Range in Alaska[13] and from New Zealand[14] and we have observed them in the Songliao basin, northeastern People's Republic of China. It appears then that the supposedly warm Cretaceous climate generated at least seasonal marine periglacial (and perhaps glacial) conditions in central Australia at palaeolatitudes of 65-70°; likewise, Northern Hemisphere ice-rafting occurred in mid-Jurassic to mid-Cretaceous times at sites located between 65°-78° palaeolatitude (Fig. 2). We suggest that because these marine areas experienced periglacial conditions, Jurassic/Cretaceous glaciation may also have occurred, although the lack of evidence could be explained in several ways[15]. Continental areas of even higher palaeolatitudes, such as Antarctica, northern USSR and northeastern Asia, Spitsbergen and Greenland, as well as Australia and New Zealand, may have had permanent glaciers, particularly at high altitudes.

Numerical modelling and sensitivity experiments on Cretaceous climates demonstrate that, even with a warm ocean, the interiors of high-latitude continents would tend to remain seasonally cold and suffer winter freezing[2,16-18]. For example, Barron and Washington[16] computed an annual temperature range from about −18 °C to +27 °C for central Australia during the mid-Cretaceous. Such strong seasonal contrast at sea level and at latitudes as high as 65°-70° imply poor heat transport from ocean to land and an intensive continentality effect. Other modelling work has produced similar results and similarly supports the notion that ice could form on at least a seasonal basis in moderately high latitudes.

Such low winter temperatures generated in the models were once considered incompatible with warm temperatures implied from high latitude Cretaceous faunas and floras[16], the principal evidence for warm polar areas. But some reinterpretations of the thermal tolerances of Cretaceous polar biotas proposed that they could well have survived long dark winters and near-freezing temperatures, perhaps even frosts[19,20]. In particular, during the early Cretaceous in southeastern Australia, forests of conifers and ferns grew in high latitudes (70-85° S), possibly associated with intermontane basins. Recent interpretations of fossil plant and fish assemblages[21-23] and dinosaurs[24] suggest that the climate was of humid, cool temperature type with marked annual seasons, including short periods of winter freezing, consistent with our sedimentary evidence.

Given that the Earth is a sphere which receives solar insolation unequally over its surface and that heat transport in both atmosphere and oceans has its physical limits, it is difficult to explain how the polar zones could ever have been warm enough to melt all traces of ice and snow there. The problem would be further exacerbated if 'normal' topographic relief to 1 to 2 km elevation existed on polar continents, producing much colder temperatures at high elevations. These factors suggest that at least

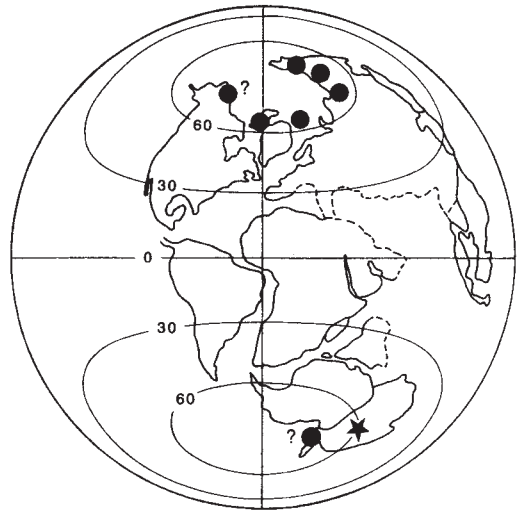


Fig. 2 Global distribution of early Cretaceous ice-rafted deposits. Star represents Eromanga Basin deposits. Reconstruction from ref. 29.

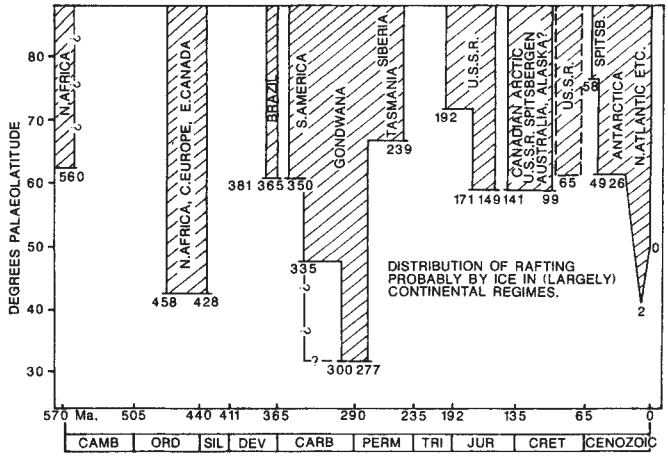


Fig. 3 Occurrence by palaeolatitude of ice-rafted deposits through the Phanerozoic. Age ranges from ref. 25 and other sources; palaeolatitudes derived from ref. 26.

seasonal ice might be expected to form in high latitudes throughout most of the Phanerozoic.

The geological literature reveals that reports of ice-rafted deposits exist for every period of the Phanerozoic Era except the Triassic. The summary of pre-Pleistocene glacial deposits published by Hambrey and Harland[25] provides most of the data. Figure 3 was constructed by plotting locations of known glacial and ice-rafted deposits on global reconstructions for the Phanerozoic and determining latitudinal distribution. Latitudes are known to within 5-10°, the standard error of measurement for the palaeomagnetic data[26]. In most cases, age of the deposits is given as a range of dates and true lengths of gaps between times of rafting are poorly known. The record of ice-rafting as seen in Fig. 3 does not completely cover the Phanerozoic; the long interval without ice-rafting in the early Phanerozoic is due, in part, to the arrangement of continents in low latitudes during this time, and the record of Palaeozoic continental glaciers (Ordovician/Silurian, late Devonian, the Carboniferous and Permian) largely reflects movement of Gondwana over the pole[27,28].

Limited geographical extent of glacial deposits in some episodes (for example, late Devonian glaciation is recorded only in South America[27]), indicates that any direct evidence of glaciation may be sparse or have gone unrecognized. It is worth

noting that the area of the globe beyond 60° (70°) latitude constitutes only 13% (6%) of the total surface area and that, for much of the Phanerozoic, landmasses covered less than half of these areas.

During times of known major glaciations in the Phanerozoic, ice-rafting extended to latitudes as low as 35°. During intervening times, ice-rafting did not reach such extremes but was limited to high latitudes. Glacial ice is likely to have occurred in association with these ice-rafted deposits and the possibility of an ice-free Earth having ever existed appears small.

This work was supported by ARGS and University of Adelaide grants. We thank N. Alley, N. M. Lemon and other colleagues for discussion.

Received 22 January; accepted 22 April 1988.

1. Hunt, B. G. *Nature* **308**, 48-51 (1984).
2. Barron, E. J. *Earth Sci. Rev.* **19**, 305-338 (1983).
3. Schmidt, P. W., Embleton, B. J. J., Cudahy, T. J. & Powell, C. McA. *Tectonics* **5**, 135-150 (1986).
4. Kurtz, D. D. & Anderson, J. B. *J. sedim. Petrol.* **49**, 1159-1170 (1979).
5. Anderson, J. B., Domack, E. W. & Kurtz, D. D. *J. Glaciol.* **25**, 387-396 (1980).
6. Emery, K. O. *J. sedim. Petrol.* **25**, 51-57 (1955).
7. Kemper, E. *Geol. Jb.* **A96**, 5-185 (1987).
8. David, T. W. E., Taylor, T. G., Woolnough, W. G. & Foxhall, H. G. *Rec. geol. Surv. N.S.W.* **8**, 161-179 (1905).
9. Kemper, E. & Schmitz, H. H. *Pap. geol. Surv. Canada* **75-1**, 109-119 (1975).
10. Epshteyn, O. G. *Int. Geol. Rev.* **20**, 49-58 (1978).
11. Embry, A. F. *CSPG-CSEG Nat. Conv. Calgary* 49-50 (1984).
12. Dalland, A. *Norsk Polarinst. Arbok.* 151-165 (1977).
13. Molenaar, C. M. *Am. Ass. Petrol. Geol. Bull.* **67**, 1066-1080 (1983).
14. Waterhouse, J. B. & Flood, P. G. in *Earth's pre-Pleistocene Glacial Record* (eds Hambrey, M. J. & Harland, W. B.) 438-446 (Cambridge University Press, 1981).
15. Crowell, J. C. & Frakes, L. A. *Am. J. Sci.* **268**, 193-224 (1970).
16. Barron, E. J. & Washington, W. M. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **40**, 103-133 (1982).
17. Barron, E. J., Thompson, S. L. & Schneider, S. H. *Science* **212**, 501-508 (1981).
18. Glancy, T. J., Barron, E. J. & Arthur, M. A. *Paleoceanography* **1**, 523-537 (1986).
19. Brouwers, E. M. *et al. Science* **237**, 1608-1610 (1987).
20. Parrish, J. M., Parrish, J. T., Hutchinson, J. H. & Spicer, R. A. *Palaios* **2**, 377 (1987).
21. Drinnan, A. N. & Chambers, T. C. *Mem. Ass. Australas. Palaeontols* **3**, 1-77 (1986).
22. Dettmann, M. E. *Mem Ass. Australas. Palaeontols* **3**, 79-110 (1986).
23. Waldman, M. *Spec. Pap. Palaeontol.* **9**, 1-124 (1971).
24. Rich, T. H. V. & Rich, P. V. *Nat. geogr. Res.* (in the press).
25. Hambrey, M. J. & Harland, W. B. *Earth's pre-Pleistocene Glacial Record* (Cambridge University Press, 1981).
26. Smith, A. G., Hurley, A. M. & Briden, J. C. *Phanerozoic Palaeocontinental Maps* (Cambridge University Press, 1981).
27. Caputo, M. V. & Crowell, J. C. *Geol. Soc. Am. Bull.* **96**, 1020-1036 (1985).
28. Frakes, L. A. *Climates Throughout Geologic Time* (Elsevier, Amsterdam, 1979).
29. Smith, A. G. & Briden, J. C. *Mesozoic and Cenozoic Palaeocontinental Maps* (Cambridge University Press, 1977).

The temperatures of oil and gas formation in the sub-surface

T. M. Quigley & A. S. Mackenzie

BP Exploration Company Ltd, Britannic House, Moor Lane, London WEC2Y 9BU, UK

Much has been written about the influence of time and temperature on the formation of oil and gas in the sub-surface[1-5]. However, a majority of previous publications have used a very simple formula to calculate the kinetics of a chemical reaction—usually referred to as the Lopatin approach[6-8]—which we believe is incorrect. Other more sophisticated schemes have generally used laboratory experiments as their primary calibrants[9]. Because reaction rates in the laboratory are at least seven orders of magnitude faster than those in nature, their relevance to sub-surface oil and gas formation is unclear, and can only be assessed provided geochemical data from deep boreholes are available[10,11]. We have calibrated a kinetic scheme that describes oil and gas formation using geological samples that have been heated under both natural and laboratory conditions. These equations predict that in the sub-surface the influence of time is not great, and that most oil is formed between 100 and 150 °C; and most gas between 150 and 220 °C.

Oil and gas are formed at elevated temperatures in the sub-surface by the transformation of the organic remains of dead organisms in fine-grained rocks[12]. Before the organic remains are raised to temperatures at which they break down into oil and gas, by the rupture of chemical bonds, most of them are solids under sub-surface conditions. These solids have very high molecular weight, and are called kerogen, which means oil-former in Greek. Kerogen can be divided into three parts based on its initial breakdown products[13]: a part that breaks down into oil called labile kerogen, a part that breaks down into gas called refractory kerogen, and a part called inert kerogen, which yields neither oil nor gas but eliminates hydrogen, oxygen, sulphur and nitrogen so as to achieve the most stable form of carbon—graphite. Labile kerogen and refractory kerogen are referred to collectively as reactive kerogen. This sub-division of kerogen is based mainly on our observations[13]; but we suspect that labile kerogen arises mainly from the lipids of algae and bacteria, and refractory kerogen from the alkyl substituents present in the lignin of higher plants. All organic matter contributes significant amounts of inert kerogen.

The progress of the individual reactions depends upon the reaction rate:

$$\frac{d[R]}{dt} = -k[R] \tag{1}$$

where $[R]$ is the concentration of the reactant, t is time and k is the reaction rate coefficient. For many chemical reactions the Arrhenius equation

$$k = A \exp(-E/RT) \tag{2}$$

defines the relationship between k and absolute temperature, T. The pre-exponential factor, A, and the activation energy, E, depend on the reaction. R is the gas constant.

The conversion of kerogen to oil and gas is assumed to consist of series of parallel first-order reactions, each of which has its own solution for equations (1) and (2). Tissot and Espitalie[9] first used this model for their Type I, II and III kerogens, each kerogen type being subdivided into six reactions. Because their model required 18 parameters to be specified for each kerogen type (six As, Es and relative initial kerogen proportions) a unique calibration is difficult[14]. Our approach is to assume that breakdown is governed by a mean pre-exponential factor ($\bar{A}$) and a gaussian distribution of activation energies, characterized by some mean ($\bar{E}$) and standard deviation (σ)[15]. Thus if we call c_{KL} the concentration of labile kerogen (normalized to its initial value) then

$$c_{KL} = \int_0^\infty f_{KL}(E) \exp\left(-\int_0^t k_{KL}(E, t)dt\right)dE \tag{3}$$

where

$$f_{KL}(E) = (2\pi)^{-1/2} \frac{\exp(-(E-\bar{E}_{KL})^2/2\sigma_{KL}^2)}{\sigma_{KL}} \tag{4}$$

and

$$k_{KL}(E, t) = \bar{A}_{KL} \exp(-E/RT(t)) \tag{5}$$

The subscript KL denotes parameters specific to labile kerogen. A similar set of equations is used to describe refractory kerogen breakdown: KL is replaced by KR everywhere in equations (3)-(5). A major advantage of this mathematical formalism is its small number of parameters (A_{KL}, E_{KL}, σ_{KL} and A_{KR}, E_{KR}, σ_{KR}). By studying the transformation of kerogens whose reactive parts are either predominantly labile or predominantly refractory we only need to optimize three parameters at a time.

Modelling the evolution of oil is more complicated. Oil is derived from labile kerogen but is cracked to gas at higher

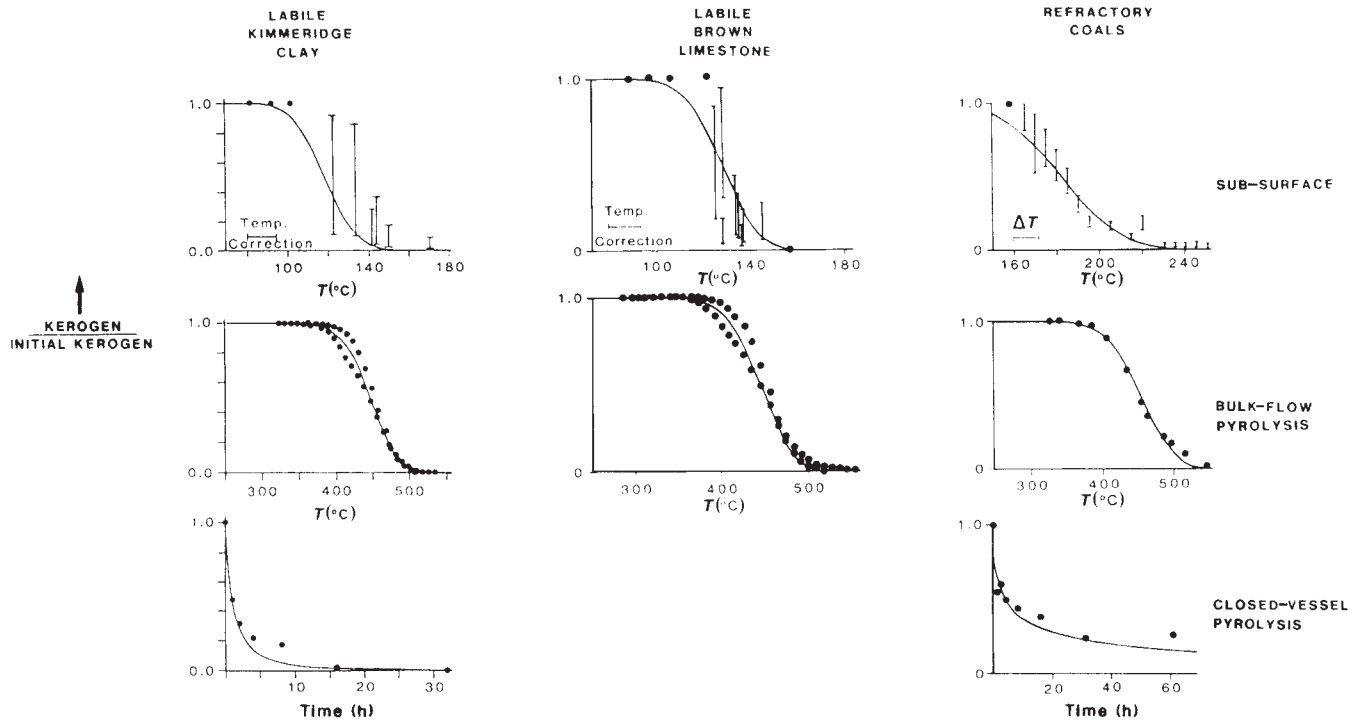


Fig. 1 Measured (data bars and dots) and calculated (solid lines) relative concentrations of reactive kerogen for three different source rocks. The reactive kerogens of the Upper Jurassic Kimmeridge Clay Formation, North Sea and the Senonian Brown Limestone Formation, Gulf of Suez are predominantly labile; the reactive kerogen of the Westphalian coals of NW Europe is predominantly refractory. The measured concentrations were obtained by techniques outlined in ref. 13. The results for kerogens heated in the sub-surface or by bulk-flow (Rock-Eval) pyrolysis are plotted as a function of maximum temperature. Consideration of thermal histories suggests that Kimmeridge Clay was heated in the sub-surface at an average rate (α) of 1 °C Myr^{-1}; Brown Limestone at an average rate of 5 °C Myr^{-1}; and Westphalian coals at an average rate of 6 °C Myr^{-1} (ref. 24). All bulk-flow pyrolysis experiments were carried out at 25 °C min^{-1}. Closed-vessel pyrolysis was performed isothermally at 370 °C: the resulting kerogen concentrations are plotted as a function of heating time. No closed-vessel pyrolysis was conducted for Brown Limestone. The data for sub-surface conditions are plotted as bars, which reflect the uncertainty when estimating original petroleum potential of mature source rocks[13]. The data for pyrolysis experiments are shown as dots. The results of two sets of bulk flow pyrolysis experiments are shown for Kimmeridge Clay and Brown Limestone to illustrate experimental reproducibility. The magnitudes of typical temperature corrections (15 °C) applied to temperature measurements down wells in the North Sea and Gulf of Suez, are shown for sub-surface data together with the estimated uncertainty (Δt) in determining the maximum temperature of coals from vitrinite reflectance (20 °C, ref. 17). The calculated concentrations (solid lines) were obtained from equations (3), (4), (5), and (9), with $E_{KL} = 208$ kJ mol^{-1}, $E_{KR} = 279$ kJ mol^{-1}, $A_{KL} = 1.58 \times 10^{13}$ s^{-1}, $A_{KR} = 1.83 \times 10^{18}$ s^{-1}, $\sigma_{KL} = 5$ kJ mol^{-1}, $\sigma_{KR} = 13$ kJ mol^{-1}.

temperatures. Although most labile kerogen breaks down to oil, a small but significant proportion of labile kerogen is also transformed directly into gas. The rate of change of concentration of oil (c_O) is

$$\frac{dc_O}{dt} = -k_O(t)c_O(t) + S_O(t) \tag{6}$$

where $S_O(t)$ is the rate of oil formation from labile kerogen. If oil also breaks down by a series of parallel first-order reactions with a gaussian distribution of activation energies, then we can integrate the appropriate version of equation (6) to yield

$$c_O(t) = \int_0^\infty dE f_O(E) \exp\left(-\int_0^t k_O(E, t')\,dt'\right) \times \int_0^t dt''\left[S_O(t'') \exp\left(\int_0^{t''} k_O(E, t''')\,dt'''\right)\right] \tag{7}$$

where

$$S_O(t) = (1-\Gamma)\int_0^\infty dE f_{KL}(E) k_{KL}(E, t) \times \exp\left(-\int_0^t k_{KL}(E, t')\,dt'\right) \tag{8}$$

Γ = mass fraction of labile kerogen that transforms into gas. (Analyses[16] of immature labile kerogen fractions by pyrolysis-gas chromatography suggest about 20% is transformed directly into molecules with less than six carbon atoms, that is, gas. Thus in the following we take $\Gamma = 0.2$.)

$f_O(E)$ and $k_O(E, t)$ are given by equations (4) and (5) respectively, with all subscripts KL replaced by the subscript O.

For the calibration of the kinetic parameters of labile kerogen two source rocks were examined in detail—Upper Jurassic Kimmeridge Clay Formation, North Sea and Senonian Brown Limestone Formation, Gulf of Suez. The reactive kerogens of these source rocks are comprised almost exclusively of labile kerogen[13]. Their thermal histories are well known, partly because their present-day sub-surface temperatures are also their measured maximum temperatures.

North-west European Westphalian coals were our refractory kerogen calibrant[13]. However, because most refractory kerogen breakdown occurs in the sub-surface at temperatures in excess of 150 °C (ref. 13)—a temperature range rarely penetrated by drilling wells—we used coals whose present temperatures are less than their maximum temperatures. The maximum temperatures of the geologically-heated samples were estimated using vitrinite reflectance measurements[17].

The thermal histories of all three can be approximately modelled using a linear heating rate α:

$$T(t) = T_0 + \alpha t \tag{9}$$

where T_0 is the initial temperature.

Using all the data shown in Fig. 1, the kinetic model described

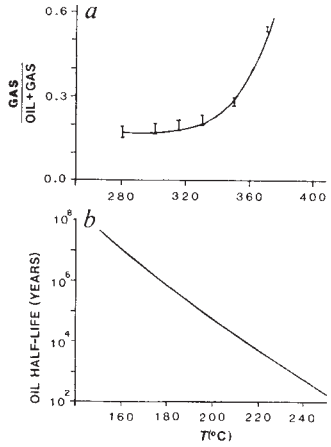


Fig. 2 Oil to gas cracking. *a*, Measured (data bars) and calculated (solid line) ratio of gas/gas + oil for isothermal pyrolysis of a Bathonian outcrop sample from NE Scotland for 3 days at the temperatures shown. The error bars reflect the uncertainty in the measurement. The reactive kerogen of the sample is predominantly labile. The solid line was calculated with equations (3)-(8) and $E_{KL} = 208$ kJ mol^{-1}, $E_O = 230$ kJ mol^{-1}, $A_{KL} = 1.58 \times 10^{13}$ s^{-1}, $A_O = 1.00 \times 10^{13}$ s^{-1}, $\sigma_{KL} = 5$ kJ mol^{-1}, and $\sigma_O = 5$ kJ mol^{-1}. *b*, Time taken to convert half a given mass of oil to gas at the temperatures shown calculated with equations (4), (5) and (7), and the kinetic parameters for oil to gas cracking reported in *a*.

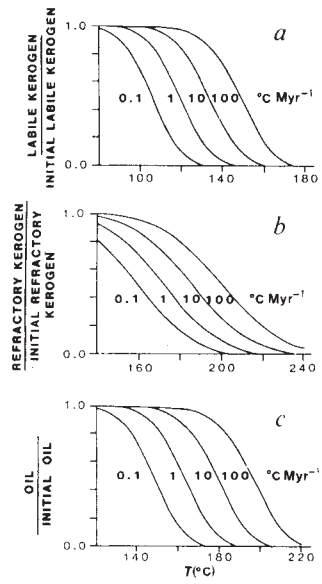


Fig. 3 Relative concentrations of *a*, labile kerogen, *b*, refractory kerogen and *c*, oil as a function of maximum temperature for the heating rates shown. The calculations use equations (2)-(9) and the kinetic parameters reported in the legends of Figs 1 and 2.

by equations (3), (4), (5) and (9), and a Gauss-Newton numerical nonlinear regression method, the labile and refractory kerogen kinetic parameters were determined. Figure 1 illustrates the performance of the model. We think the kinetic parameters of the model are not greatly affected by differences in mineral matrix because labile kerogen breakdown in a clay and limestone can be described using the same kinetic parameters.

There are no well constrained geological observations of oil to gas cracking. Therefore we must use primarily laboratory experiments to calculate A_O, E_O and σ_O; with empirical observations of the temperatures above which oil is not found in Nature as secondary calibrants.

Closed-vessel pyrolysis experiments were conducted using Bathonian outcrop source rock samples from North-east Scotland whose reactive kerogen is predominantly labile. Sealed platinum capsules were used, which at the end of the experiment were burst in the injector of a gas chromatograph to analyse the relative concentrations of gas (molecules with less than six carbon atoms) and oil (molecules with more than five carbon atoms). The results are shown in Fig. 2*a*. The oil cracking kinetic parameters were obtained by calibrating equations (3) to (8) using the data shown in Fig. 2*a*. Similar series of experiments and calculations were performed using other source rocks of varying mineralogy, from aluminosilicate through diatomite to carbonate. No significant variations from the results shown in Fig. 2*a* were observed: we therefore use the Bathonian outcrop results as a typical example of the oil to gas cracking reaction.

The time taken for half of a given mass of oil to be converted to gas at a given temperature—the oil half-life—can be calculated using a modified version of equation (7). Figure 2*b* predicts an oil half-life of less than 1 Myr (a short time on a geological timescale) for temperatures exceeding 170 °C. The experience of the oil exploration industry supports this because oil does not exist in significant amounts above this temperature in the sub-surface, but often occurs at temperatures as hot as 160 °C (ref. 18).

Our calibrated kinetic model allows us to explore the importance of temperature and time for oil and gas formation. Fig. 3 displays the calculated extent of labile and refractory kerogen breakdown, and oil cracking to gas, as a function of maximum temperature for a range of geologically reasonable heating rates. The faster the rate the greater the temperatures of the three processes. But in practice very few sedimentary sequences unaffected by igneous intrusions are heated at rates outside the range 1-10 °C Myr^{-1} (ref. 19). Within this range, Fig. 3 shows that the difference in temperature associated with a particular kerogen concentration of a given kerogen type (labile or refractory) is less than 15 °C. A similar observation can be made for the oil cracking reaction. This is often within the accuracy to which we can measure temperatures in wells[20,21] or reconstruct the thermal history of a sediment[22]. So for organic matter buried in sedimentary basins the effect of time can often be neglected, and the temperature ranges apparent from Fig. 3 can be used directly: oil (and gas) formation from labile kerogen breakdown, 100-150 °C; gas formation from refractory kerogen breakdown, 150-220 °C; oil to gas cracking, 150-190 °C. Our own experience from many sedimentary basins around the world—North Sea, Gulf of Suez, Norway, Hungary, France, Western Canada, offshore California, Japan, Cameroon—convinces us that these temperature guidelines provide a valuable petroleum exploration tool to aid explorers assess the overall prospectivity of areas of interest rapidly.

The oil to gas cracking reaction must consume additional hydrogen if it goes to completion. Hydrogen could be supplied from the rearrangement of aromatic parts of inert kerogen to graphite-like structures; clay mineral dewatering; or even water[23]. We assume hydrogen is available in excess; if it is not, then some of the oil molecules must form more condensed and aromatic structures in order to release extra hydrogen. If this is occurring in our closed-vessel pyrolysis experiments the temperature range of oil to gas cracking would be narrower than the prediction of our model for the sub-surface but the onset of the reaction would still occur at 150 °C.

Finally, we have used our calibrated model to test the validity of the so-called Lopatin approach[6-8]. This approach derives from a rule-of-thumb used in the chemical laboratory or in the photographic development laboratory and states that the rate of petroleum-forming reactions double for every 10 °C rise in temperature. We find that the Lopatin approach overestimates the effect of time and underestimates the effect of temperature on source rock maturation[17].

We thank BP for permission to publish, and J. R. Gray for providing the results of his experiments.

Received 29 October 1987; accepted 19 April 1988.

1. Phillipi, G. T. *Geochim. cosmochim. Acta* **29**, 1021-1049 (1965).
2. Hood, A., Gutjahr, C. C. M. & Heacock, R. L. *Bull. Am. Assoc. Petrol. Geol.* **59**, 986-996. (1975).
3. Connan, J. *Bull. Am. Assoc. Petrol. Geol.* **59**, 2516-2521 (1974).
4. Price, L. C. *J. Petrol Geol.* **6**, 5-38 (1983).
5. Siever, R. *Bull. Am. Assoc. Petrol. Geol.* **67**, 684-691 (1983).
6. Lopatin, N. V. *Izvestiya Akademii Nauk USSR, Seriya Geologicheskaya* **3**, 95-106 (1971).
7. Waples, D. W. *Bull. Am. Assoc. Petrol. Geol.* **64**, 916-926 (1980).
8. Ejedawe, J. E. & Coker, S. J. L. *Bull. Am. Assoc. Petrol. Geol.* **68**, 1024-1028 (1984).
9. Tissot, B. & Espitalie, J. *Rev. Inst. Fr. Petrol.* **30**, 743-777 (1975).
10. Ungerer, P. & Pelet, R. *Nature* **327**, 52-54 (1987).
11. Sweeney, J. J., Burham, A. K. & Braun, R. L. in *Thermal Modeling in Sedimentary Basins* (ed. Burrus, J.) 547-561 (Technip, Paris, 1986).
12. Tissot, B. P. & Welte, D. H. *Petroleum Formation and Occurrence 2nd edn* (Springer, Berlin, 1984).
13. Cooles, G. P., Mackenzie, A. S. & Quigley, T. M. *Org. Geochem.* **10**, 235-245 (1986).
14. Ungerer, P. in *Thermal Phenomena in Sedimentary Basins* (ed. Durand, B.) 235-246 (Technip, Paris, 1984).
15. Pitt, G. J. *Fuel* **41**, 267-274 (1962).
16. Mann, A. L., Goodwin, N. S. & Lowe, S. in *16th Convention of Indonesian Petroleum Association* 241-258 (1987).
17. Quigley, T. M., Mackenzie, A. S. & Gray, J. R. in *Migration of Hydrocarbons in Sedimentary Basins* (ed. Doligez, B.) 649-665 (Technip, Paris, 1987).
18. Price, L. C. *Chem. Geol.* **28**, 1-30 (1980).
19. Gretener, P. E. & Curtis, C. D. *Bull. Am. Assoc. Petrol. Geol.* **66**, 1124-1149 (1982).
20. Nwachukwu, S. O. *Bull. Am. Assoc. Petrol. Geol.* **60**, 1073-1077 (1976).
21. Luheshi, M. N. *Geophys J-R. astr. Soc.* **74**, 747-766 (1983).
22. Guidish, T. M. *et al. Bull. Am. Assoc. Petrol. Geol.* **69**, 92-105 (1985).
23. Hoering, T. C. *Org. Geochem.* **5**, 267-278 (1984).
24. Juntgen, H. & Klein, J. *Erdol und Kohle-Erdgas-Petrochemie* **28**, 65-73 (1975).

The discovery of Santorini Minoan tephra in western Turkey

Donald G. Sullivan

Department of Geography, University of Denver, Denver, Colorado 80208, USA

Dispersal of volcanic ash from the violent Bronze Age (Minoan) eruption of the Santorini volcano in the southern Aegean has been the subject of much research[1-6]. A sediment core taken from a small mountain lake in western Turkey contains a layer of tephra from the Minoan eruption. The Santorini origin is established by refractive index, major glass chemistry, and stratigraphic position. The lake is located north of the recognized dispersal area for the tephra. This deposit provides evidence for dispersal of the Minoan tephra well to the northeast of Santorini, and implies that the west coast of Asia Minor must have experienced heavy tephra fallout from the Bronze Age eruption. The Minoan tephra has not previously been identified from sites in Turkey.

A 12-cm-thick layer of Minoan tephra was discovered in a sediment core recovered from Gölcük, a small lake (elevation 1,050 m) in the Bozdağ Range, 90 km east of the Aegean port of Izmir, and about 320 km northeast of Santorini (Fig. 1). A description of the sediments is shown in Fig. 2.

The tephra was found in a peat and peaty mud stratigraphic unit, implying that the layer was deposited during a shallow stand of the lake. The contact between the tephra and the underlying peat is sharp, indicating that the tephra was deposited rapidly. Plant roots in growth positions penetrate the layer, but there are no lacustrine muds mixed with the tephra, nor any indication of grading, suggesting that the tephra was not subject to sediment ponding or secondary depositional processes, as has been found in some of the tephra deposits recovered in Aegean deep-sea cores[2,7].

The contact between the tephra and the overlying peat is gradational, implying bioturbation. Watkins *et al.*[2] found that upward mixing of sediments resulting from bioturbation could account for a decrease in the thickness of tephra layers in cores from the southern Aegean. Other workers[8,9] have reported similar findings. In addition, Watkins *et al.* reported that post-depositional compaction may reduce tephra thickness in deep-sea cores by ~50%. If conditions in a lacustrine environment are comparable, the 12-cm-thick tephra layer in the Gölcük sediment core represents the equivalent of at least 24 cm of freshly fallen tephra on land.

Fig. 1 Map of the Aegean Sea and western Turkey showing sites mentioned in the text. The newly discovered Minoan tephra layer was recovered from Gölcük, a small lake in the Bozdağ Range, ~320 km northeast of Santorini.

The Gölcük tephra layer is 30 cm below a peat deposit yielding a radiocarbon date of 3,110 ± 160 BP. A radiocarbon date of 7,400 ± 120 BP on peat from below the tephra indicates that the tephra layer is of mid- to late-Holocene age. The Minoan tephra is the only widely dispersed Holocene tephra layer in Aegean deep-sea cores (where it is referred to as the Z-2 tephra)[3]. Using the new high-precision radiocarbon calibration curves derived by Pearson and Stuiver[10], the 3,110 BP radiocarbon date for the peat overlying the tephra calibrates to a calendric date of ~1420 BC.

An archaeological date for the Bronze Age eruption of Santorini is a matter of some controversy, with dates from the mid-seventeenth to the early fifteenth century BC suggested[11,12]; proxy evidence from tree rings and ice cores implies a mid-seventeenth century BC date for the eruption[13-15].

The refractive index of glass shards in the Gölcük tephra is 1.508 ± 0.002. Most authors report an index of refraction for the Minoan tephra of 1.508 to 1.510 (refs 1-3, 5, 6). In deep-sea cores from the southern Aegean Sea, the only tephra with a similar refractive index is the Pleistocene V-1 tephra, probably

Table 1 Major element glass chemistry determined by electron microprobe analysis

SiO_2	Al_2O_3	FeO*	MgO	CaO	K_2O	Na_2O†	TiO_2
73.36	13.94	2.09	0.30	1.41	3.13	4.22	0.30

Values averaged for 16 sample analyses.
* Total iron expressed as FeO.
† Some Na volatilization noted[17]; Na_2O value derived from 1-second counts.

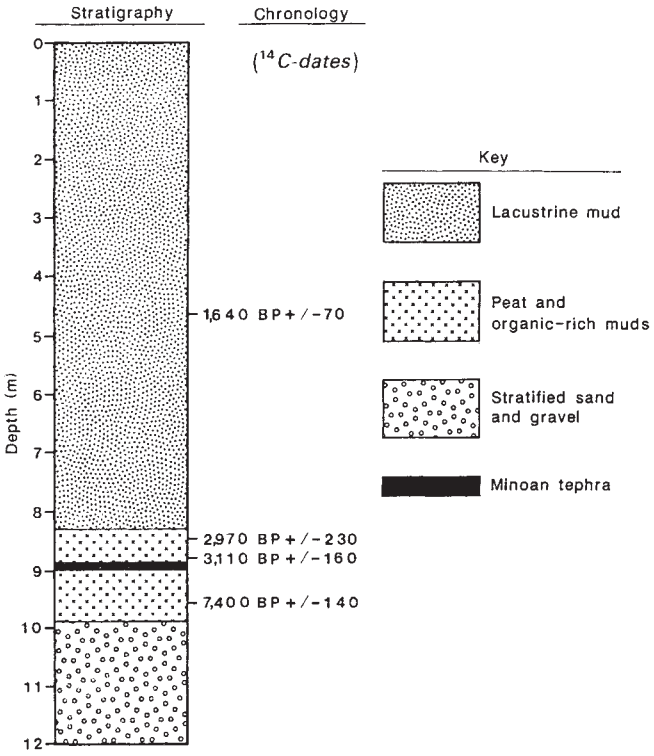


Fig. 2 Stratigraphy of the Gölcük sediment core. Minoan tephra layer is bracketed by radiocarbon dates of 3,110 BP (Beta Analytic 6,102) and 7,400 BP (Beta 12,543).

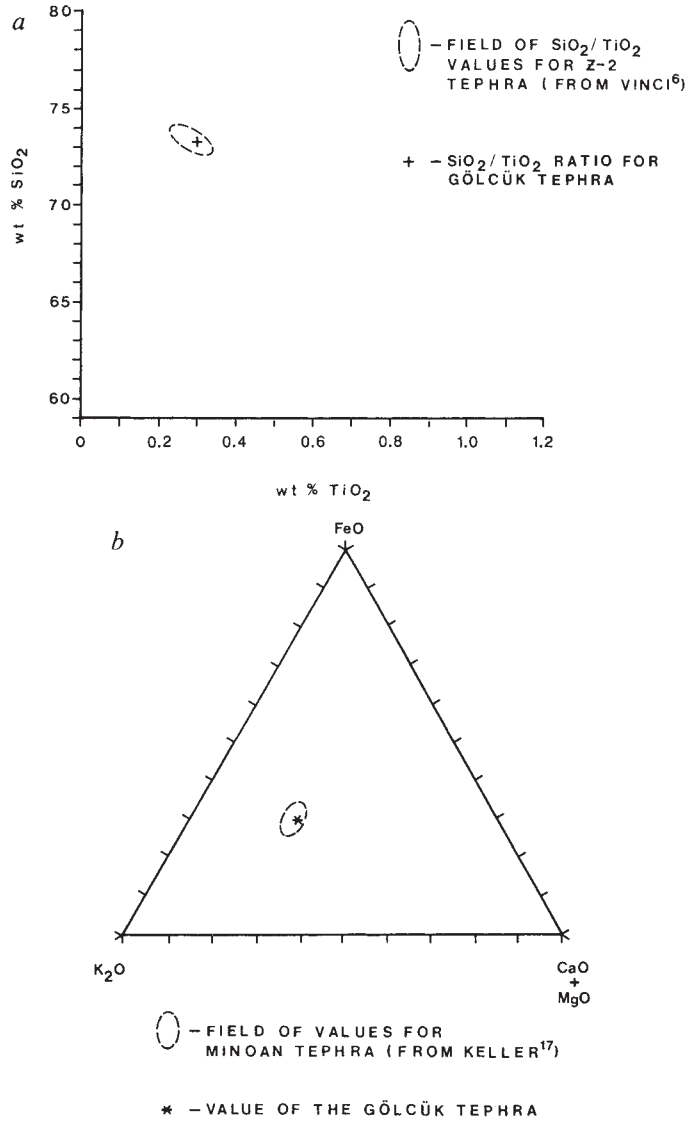


Fig. 3 *a*, SiO_2/TiO_2 ratios for Gölcük tephra and Z-2 tephra (after Vinci[6]). Gölcük ratio falls within the range of values for Z-2 tephra. *b*, Ternary plot of $FeO-K_2O-CaO+MgO$ for Gölcük tephra and Minoan tephra samples (after Keller[17]). Gölcük tephra values fall within the range of values for the Minoan tephra.

erupted from one of the eastern Hellenic arc volcanoes around 160,000 years ago; the Pleistocene Y-2 and Y-4 tephras from Santorini have refractive indices of 1.514 and 1.518, respectively[3].

The major element glass chemistry of the Minoan tephra is distinctive, and provides unambiguous identification[2,6]. Table 1 shows the results of electron microprobe analysis on glass shards from the Gölcük tephra layer. The major element glass chemistry conforms with the published results of microprobe analyses of the Minoan tephra[2,5,6,16,17].

Vinci[6] reported that SiO_2/TiO_2 ratios may be used to differentiate between Hellenic arc tephras; and Keller[17], and others[2,16] used ternary plots of $FeO-K_2O-CaO+MgO$ to characterize the Minoan tephra. Figure 3*a* shows the SiO_2/TiO_2 ratios published by Vinci for the Z-2 (Minoan) tephra, and the same ratio for the Gölcük tephra; Fig. 3*b* shows a plot of $FeO-K_2O-CaO+MgO$ for the Gölcük tephra, compared with the field of values for the Minoan tephra given in Keller[17].

The radiocarbon dating, major element glass chemistry, SiO_2/TiO_2 ratios and plots of $FeO-K_2O-CaO+MgO$ leave little doubt that the Gölcük tephra layer is from the Minoan eruption of Santorini.

The magnitude and dispersal direction of the tephra fall from the paroxysmal Bronze Age eruption of Santorini have been a focus of research since Marinatos[18] linked the eruption of Santorini with the decline of the Minoan settlements on Crete. Since the late 1940s, the presence of Minoan tephra in deep-sea cores from the southern Aegean and eastern Mediterranean Seas has been reported[1-3,6,16,19]. These reports serve as the main basis for the dispersal maps published to date.

Ninkovich and Heezen[1] published the first dispersal maps for the Minoan tephra, indicating a southerly to southeasterly axis of dispersal. Watkins *et al.*[2] published isopach maps showing the thickness of Minoan tephra in sediment cores recovered from south and southeast of Santorini. On the basis of their maps, they inferred a somewhat more easterly dispersal axis for the Minoan tephra than that shown by Ninkovich and Heezen[1], but still generally to the southeast, with the fallout axis passing over the island of Karpathos. The authors noted that a more easterly axis could not be ruled out, and tephra dispersal to the north could not be determined due to the lack of cores north of Santorini. A later isopach and distribution map[19] shows a nearly easterly dispersal for the tephra.

Additional information on the dispersal of the tephra has come from terrestrial deposits. Keller[5] has reported finding 30 cm of Minoan tephra at the east end of Kos (about 190 km east-northeast of Santorini), and Doumas and Papazoglou[20] described a 10-cm-thick layer of Minoan tephra from the archaeological site at Trianda on Rhodes (~225 km east of Santorini). More recently, at least 0.5 m of Minoan tephra has been discovered in a construction trench at the Rhodes airport (C. Doumas, personal communication). Taken with the 12-cm-thick layer at Gölcük (about 320 km to the northeast), these findings argue that a principally southeasterly dispersal of the Minoan tephra is no longer tenable.

These findings also suggest that Minoan tephra deposits from terrestrial and lacustrine environments may give a better indication of atmospheric distribution than do deep-sea sediments,

the latter subject to dispersal by the predominantly southeasterly surface and intermediate water currents in the Aegean basin[21-23].

Watkins *et al.*[2] calculated a dense rock equivalent volume of 13 km^3 of tephra within their 0.5 cm isopach. They noted that their estimate was very much a minimum as the volume of tephra outside the 0.5 cm isopach could be substantial, and because there were insufficient data available north and east of Santorini. They further noted that there was a substantial discrepancy between their calculated dense rock equivalent volume (13 km^3) and the volume of the caldera at Santorini (60 km^3, ref. 1). This discrepency led Heiken and McCoy[24] to argue that much of the volume of the caldera at Santorini must have resulted from an earlier eruption and caldera collapse.

Although it would be premature to attempt to re-calculate the total volume of Minoan tephra based on the data presented here, it is clear that a dense rock equivalent volume considerably greater than 13 km^3 can now be expected. That such a large volume of tephra was ejected lends support to the interpretations of dendrologic evidence for the impact of the Santorini eruption on global climate by LaMarche and Hirschboeck[13] and Baillie and Munro[15].

The implications of the discovery of Minoan tephra at Gölcük are several. Ninkovich and Heezen[1], assuming a southeasterly dispersal of the tephra, suggested that the eruption occurred during the summer months, when the prevailing winds blow from the northwest. The presence of substantial Minoan tephra deposits on Rhodes, on Kos and in western Turkey indicates that the major axis of dispersal must have been to the east and/or northeast of Santorini, and implies that the eruption took place when winds would have been blowing from the west or southwest. Assuming that atmospheric circulation ~3,500 years ago was generally similar to the present, such winds would be most common during the winter and spring months, when jetstream flow and central basin depressions might drive the tephra to the northeast[25].

Keller[5] argued that the presence of the thick deposit of tephra on Kos indicated that western Anatolia must have suffered severely from Minoan tephra fall-out. The minimum fresh-fall equivalent thickness of the tephra layer from Gölcük (~24 cm) confirms Keller's suggestion. The extent to which the Minoan tephra might have affected Bronze Age settlements is difficult to evaluate. Reports by Blong[26,27] and Thorarinsson[28] indicate that tephra fallout of the magnitude shown in the Gölcük core may have been sufficient to damage structures and severely disrupt agriculture.

Finally, the presence of the tephra layer in Turkey provides a stratigraphic and chronologic marker that will be useful in correlating Holocene palynological studies[29] of lake core sediments from western Asia Minor.

Future work will focus on recovery of Minoan tephra from other lakes in Turkey, improved dating of sediments contemporary with the tephra layer, and palynological evidence of short-term effects of the tephra fallout on vegetation in western Turkey.

The author acknowledges the assistance of Theodore Chenoweth in identifying the tephra layer; Professor Lionel Weiss for the refractive index determination; John Donovan for electron microprobe analysis; and Kaya Bengisu for his assistance in obtaining permission to work at Gölcük. Figures 1-3 were drafted by Eric Schwab, Department of Geography, University of Denver. This research · s supported by the Harvard-Cornell Archaeological Expedition at Sardis, and by a National Science Foundation grant to the Sardis Urban Survey Project.

Received 23 November 1987; accepted 14 April 1988.

1. Ninkovich, D. & Heezen, B. C. *Colston Res. Papers* **17**, 413-452 (1965).
2. Watkins, N. D. *et al. Nature* **271**, 122-126 (1978).
3. Keller, J., Ryan, W. B. F., Ninkovich, D. & Altherr, R. *Geol. Soc. Am. Bull.* **89**, 591-604 (1978).
4. McCoy, F. W. in *Thera and the Aegean World II* (ed. Doumas, C.) 57-78 (Aegean World, London, 1980).
5. Keller, J. in *Thera and the Aegean World II* (ed. Doumas, C.) 49-56 (Aegean World, London, 1980).
6. Vinci, A. *Mar. Geol.* **64**, 143-155 (1985).
7. McCoy, F. Jr in *Tephra Studies* (eds Sparks, R. S. J. & Self, S.) 245-254 (Reidel, Dordrecht, 1981).
8. Ninkovich, D. & Ruddiman, W. *Proc. X. INQUA Congress* (International Quaternary Assoc., 1977).
9. Ninkovich, D., Shackelton, N. J., Abdel-Monem, A. A., Obradovich, J. D. & Izett, G. *Nature* **276**, 574-577 (1978).
10. Pearson, G. W. & Stuiver, M. *Radiocarbon* **28**, 839-862 (1986).
11. Warren, P. *Nature* **308**, 492-493 (1984).
12. Betancourt, P. P. *Archaeometry* **29**, 45-59 (1987).
13. LaMarche, V. C. & Hirschboeck, K. K. *Nature* **307**, 121-126 (1984).
14. Hammer, C. U., Clausen, H. B., Friedrich, W. L. & Tauber, H. *Nature* **328**, 517-519 (1987).
15. Baillie, M. G. L. & Munro, M. A. R. *Nature* **332**, 344-346 (1988).
16. Federman, A. N. & Carey, S. N. *Quat. Res.* **13**, 160-171 (1980).
17. Keller, J. in *Tephra Studies* (eds Sparks, R. S. J. & Self, S.) 227-244 (Reidel, Dordrecht, 1981).
18. Marinatos, S. *Antiquity* **13**, 425-439 (1939).
19. Mellis, O. *Deep-Sea Res.* **2**, 89-92 (1954).
20. Doumas, C. & Papazoglou, L. *Nature* **287**, 322-324 (1984).
21. Lacombe, H., & Tchernia, P. in *The Mediterranean Sea* (ed. Stanley, D. J.) 25-36 (Dowden, Hutchinson & Ross, Stroudsburg, Pennsylvania, 1972).
22. Miller, A. R. in *The Mediterranean Sea* (ed. Stanley, D. J.) 37-42 (Dowden, Hutchinson & Ross, Stroudsburg, Pennsylvania, 1972).
23. Venkatarathnam, K., Biscaye, P. E. & Ryan, W. B. F. in *The Mediterranean Sea* (ed. Stanley, D. J.) 455-469 (Dowden, Hutchinson & Ross, Stroudsburg, Pennsylvania, 1972).
24. Heiken, G. & McCoy, F. Jr *J. geophys. Res.* **89**, 8441-8462 (1984).
25. Wigley, T. M. L. & Farmer, G. *B.A.R. Int. Ser.* **133**, (eds Bintliff, J. L. & van Zeist, W.) 3-37 (British Archaeological Reports, Oxford, 1982).
26. Blong, R. J. in *Thera and the Aegean World II* (ed. Doumas, C.) 217-226 (Aegean World, London, 1980).
27. Blong, R. J. in *Tephra Studies* (eds Sparks, R. S. J. & Self, S.) 405-420 (Reidel, Dordrecht, 1981).
28. Thorarinsson, S. *Acta Int. Sci. Congr. Volcano Thera* 213-236 (Greek Archaeological Science, Athens, 1971).
29. Buckland, P. C., Foster, P., Perry, D. W. & Savory, D. in *Tephra Studies* (eds Sparks, R. S. J. & Self, S.) 381-289 (Reidel, Dordrecht, 1981).

Extinct pygmy hippopotamus and early man in Cyprus

Alan H. Simmons

Desert Research Institute, University of Nevada System, POB 60220, Reno, Nevada 89506, USA

Little evidence exists for human occupation of the Mediterranean Islands before ~9,000 yr before present (BP)[1], and extinct Pleistocene fauna rarely have been associated with archaeological materials from the region. Intriguing claims, however, have been presented for the early occupation of Mallorca[2] and Sardinia[3]. Although it has been suggested that Cyprus may have been inhabited as early as the Palaeolithic, this has not been substantiated and the consensus of opinion is that the initial inhabitation was by Neolithic peoples ~9,000 yr ago[4,5]. But new evidence from Akrotiri Aetokremnos ('Eagle's Cliff' or 'Site E') suggests an earlier occupation. This site also contains evidence for the association of cultural remains with extinct pygmy hippopotamus (*Phanourios minutus*[6]). Although beds containing bones of the pygmy hippopotamus are known from Cyprus[7,8], this species is generally believed to have gone extinct before man's presence on the island[9]. From my analysis, I tentatively conclude that Site E is the earliest archaeological site in Cyprus and that the Island was occupied by approximately 10,000 yr BP. Finally I suggest that the extinction of the pygmy hippopotamus on Cyprus may have been accelerated by the presence of man on the island.

Site E was discovered by a fossil hunter in 1980 and is located on the Akrotiri Peninsula in southern Cyprus (Fig. 1). It sits on the narrow talus of a precipitous cliff some 60 m above the Mediterranean. The site is a collapsed rockshelter, portions of which have eroded into the sea. Surface indications of cultural residues are scant. A few chipped stone artefacts are present, but the most suggestive evidence is a layer, 10-15 cm thick and ~2 m long, of burned and cracked marine shell capping another layer of bone in ashy soil. This deposit resembles a cultural refuse midden. It was exposed by erosion, and its extent is unknown. Spread across the surface of the site is bone, some burned, the majority of which has been identified by David

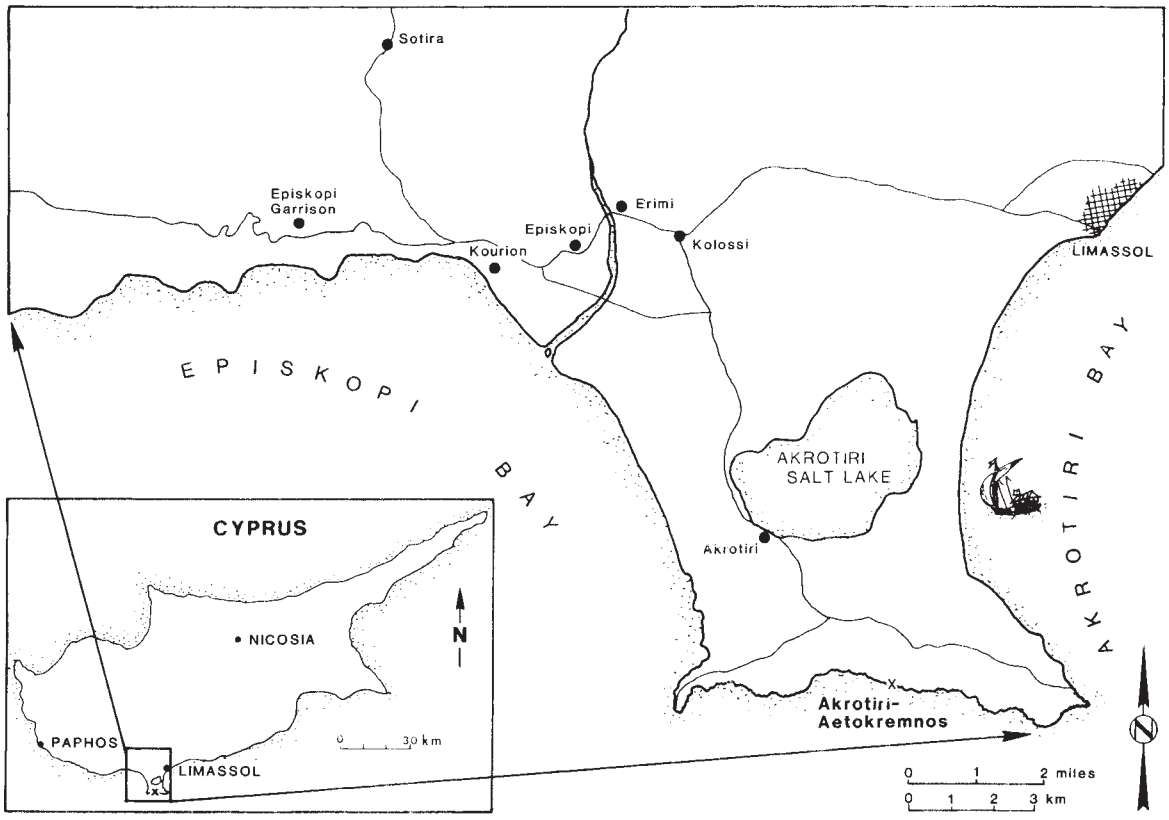


Fig. 1 Map of the Akrotiri Peninsula, Cyprus, showing the location of Akrotiri-Aetokremnos.

Reese (Field Museum of Natural History), an expert in extinct Mediterranean fauna[10,11], as pygmy hippopotamus.

In 1984, exposed samples of both shell and bone were submitted for radiocarbon dating (Table 1). Although these dates suggested considerable antiquity, there were some problems. If the shells dated to ~10,200 BP represent a midden, Site E is some 1,200 years older than any other site in Cyprus. The bone dates, however, were equivocal in confirming the site's antiquity. If this evidence could be verified, Site E would be significant from two perspectives. First, it would represent the earliest archaeological site in Cyprus and, indeed, one of the oldest occupations in the Mediterranean. Second, if a direct association of pygmy hippopotamus with cultural remains could be demonstrated, the extinction of these animals would require re-evaluation.

During July of 1987 a short (2 week) excavation was conducted. The primary objective was to determine if Site E was cultural. The results strongly strengthen the case for it being a cultural phenomenon. Exposed bone was systematically collected and three square meters were excavated, two at the exposed shell layer and one ~1 m away. This revealed far more extensive intact materials than anticipated. The abundance of bone was greater than expected; ~120 kg of material was retrieved, most of it pygmy hippopotamus. Much of the bone lay on the exposed floor of the shelter and is badly weathered, but the large amount that was buried is remarkably fresh in appearance. A wide range of both adult and juvenile pygmy hippopotamus cranial and post-cranial elements were recovered. These include teeth, mandibles and complete long bones, clearly representing several individuals.

The most dramatic result of the excavation was the documentation of ~75 cm of intact deposit. This appears to continue to the back of the shelter, where it may be thicker. The upper 10–15 cm is the shell layer exposed by erosion. Immediately underlying this is a midden-like deposit of non-fossilized bone in a matrix of ashy soil and shell (Fig. 2). There appears to be a disproportionate number of cranial elements, including one intact skull; few pieces are articulated and several are partially burned. This distribution suggests differential discard of body parts rather than a natural 'die site' or deposition by natural causes. Confirmation of this patterning will be addressed by future analysis.

Fig. 2 Part of the excavated section at Akrotiri-Aetokremnos, after excavation to the rockshelter's bedrock floor. Note shell layer (top) and ashy deposit with bone. Scale bar, 0.5 m.

A feature was also exposed in the unit excavated away from the shell layer, outside of the midden. This consists of a cone of dark matrix with a burned red base. The feature was associated with bone and chipped stone.

Table 1 Radiocarbon dates from Akrotiri-Aetokremnos

Date (BP)	Laboratory	Material	Provenience/comments
3,700 ± 60	Pta-3435	Bone collagen	Surface
6,310 ± 160	Beta-3412	Charred bone	Surface
8,330 ± 100	Pta-3281	Charred bone	Surface
9,040 ± 160*	TX5976	Bone apatite	N95E88, Level 1/top of Level 2
9,100 ± 790†	ISGS 1743	Total organics from bone	N95E88, Level 1/top of Level 2
9,240 ± 420‡	TX5833C	Humins fraction§	Feature 1, ca. 60 cm below present ground surface; N98E88/87, Level 4
9,250 ± 150	Pta-3128	Charred bone	Partially exposed beneath shell layer
9,490 ± 120	TX5833A	Bulk organic carbon	Feature 1
10,120 ± 110‖	Beta-22811	Shell	N98E88, Level 1; in dark midden-like matrix containing bone and artefacts below exposed shell layer
10,150 ± 60‖	SMU-1991	Shell	N98E88/87, Level 4; outside of midden area; associated with bone, artefacts, and Feature 1
10,150 ± 130	TX5833B	Humic acid fraction¶	Feature 1
10,280 ± 100‖**	Pta-3322	Shell	Exposed shell layer
10,310 ± 100‖**	Pta-3112	Shell	Exposed shell layer

All of the Pta dates, as well as Beta-3412, were obtained before the 1987 test excavations, thus exact provenience information is not known. The exposed surface specimens are likely to have been contaminated due to their long exposure; this may well account for the wide range in dates. All of the bone dates presented here are from pygmy hippopotamus. All dates except Beta-3412 are C-13 adjusted.

* This date is from seven pygmy hippopotamus bone fragments from the same excavation unit and level.

† As with TX5976, this date also is from seven pygmy hippopotamus bone fragments. These are from the same excavation unit and level as were the TX5976 specimens. The large sigma is due to the small amount of total organics, including collagen, present in the sample (0.2 grams from a total sample weight of 400 grams).

‡ All three TX5833 dates are from the soil matrix of Feature 1.

§ The humins fraction is a very small sample consisting of the insoluble humate fraction.

‖ There is some question as to how many years need to be subtracted from shell dates to account for the 'reservoir effect', which is an estimated correction for surface ocean water contamination; 400 years frequently is used[14]; Struiver's marine calibration curve allows calibrations up to ~8,500 BP. The local ^{14}C correction, which compensates for upwelling and evaporation effects, has only one point listed for the Mediterranean Sea, off the coast of Algeria. With this information, a calibrated age of 690 years older than the ^{14}C age is calculated by Struiver's calibration program and speculative studies for ages beyond the established calibration curve indicate corrections of the same size or larger of 10,000 years BP (H. Hass, Southern Methodist University Radiocarbon Laboratory, personal communication; M. Tamers, Beta Analytic, personal communication). Accordingly, all of the shell dates presented here have had 690 years subtracted from the ^{14}C age.

¶ The greater age of this sample compared to the other TX5833 specimens may be due to the presence of old carbon in the humic acid fraction.

** Both of these dates are from the same specimen; inner and outer fractions.

Table 2 Summary of lithic artefacts recovered from Akrotiri-Aetokremnos

			Excavation†				
	Surface*	level 1	Level 2	Level 3	Level 4		Total
Type	*N*	*N*	*N*	*N*	*N*	*N*	%
Tools							
side scraper	1		1			2	3.1
end scraper	2		1			3	4.7
scraper plane	1					1	1.6
retouched cortical flake				1		1	1.6
retouched secondary flake	1					1	1.6
Debitage							
cortical flake	1				1	2	3.1
secondary flake	18	1		4	3	26	40.6
blade	7	2				9	14.6
bladelet					1	1	1.6
Microflakes	1	1	1	2	3	8	12.5
Debris	2				1	3	4.7
Cores							
single platform, flake	1					1	1.6
Hammerstones	2					2	3.1
Groundstone							
quern fragments	3					3	4.7
incised weight(?)	1					1	1.6
Totals	41	4	3	7	9	64	100.2

* From entire site, does not include five artefacts (two secondary flakes, three debris) located downslope at shore east of site.
† Primarily from three 1 × 1-m units.

A thorough survey of the area from the cliff top down to the site revealed no lithic artefacts. But chipped stone artefacts and one shell bead were recovered during excavation. The presence of chipped stone in buried and sealed deposits, coupled with the presence of such artefacts on the surface of the site and below it (Table 2), but not above it, indicates *in situ* cultural remains.

Site E's chronology, while not entirely resolved, is becoming clearer. Bone, shell, and soil recovered during the test excavations have yielded several additional radiocarbon dates (Table 1), all of which support the antiquity of the site.

One reason for the disparate age range from the previously dated bone may well be that the samples were taken from exposed surface contexts. Exposure for several centuries could have resulted in contamination[12,13]. If one rejects the surface specimens as contaminated, a remarkably consistent pattern

emerges, despite the difficulties of dating both shell and bone. Ironically, Site E is now one of the best dated archaeological sites in Cyprus.

Based on the test excavations, I tentatively conclude that both points of significance mentioned earlier are correct: Site E is the earliest site in Cyprus and it appears that pygmy hippopotami were indeed hunted. Whether or not these animals became extinct because of human intervention cannot be determined on the basis of limited excavations at one site, but the implications are intriguing. It is essential that additional research be undertaken before the site completely erodes into the Mediterranean Sea.

This project could not have been undertaken without the volunteer effort of several very motivated professionals, the encouragement of Dr V Karageorghis and the Cypriot Department of Antiquities, the support of Dr S Swiny and the Cyprus American Archaeological Research Institute, and the assistance of the Western Sovereign Base Archaeological Society (Royal Air Force, Akrotiri). Partial funding was provided by the Desert Research Institute through a grant from the National Science Foundation's Experimental Program to Stimulate Competitive Research (EPSCoR).

Received 8 January; accepted 29 April 1988.

1. Cherry, J. *Proc. Prehist. Soc.* **47**, 41-68 (1981).
2. Burleigh, R. & J. Clutton-Brock *J. Archaeol. Sci.* **7**, 385-388 (1980).
3. Sondaar, P., Sanges, M., Kotsakis, T. & deBoer, P. *Geobios* **19**, 17-25 (1986).
4. Stanley-Price, N. *Levant* **9**, 66-89 (1977).
5. Stanley-Price, N. *World Archaeol.* **9**, 27-41 (1977).
6. Faure, M., Guerin, C. & Sondaar, P. *Actes Symp. Paleont. Cuvier, Montebeliard* 157-183 (1983).
7. Boekschoten, G. & Sondaar, P. *Konikl. Nederl. Akad. van Weter* **75**, 306-338 (1972).
8. Bate, D. *Geol. Mag.* **5**, 324-325 (1904).
9. Davis, S. *The Archaeology of Animals* (Yale, New Haven, 1987).
10. Reese, D. *Earth Sci.* **28**, 63-69 (1975).
11. Reese, D. *Expedition* **17**, 26-30 (1975).
12. Vogel, J., Fuls, A. & Visser, E. *Radiocarbon* **28**, 1133-1172 (1986).
13. Vogel, J. & Waterbolk, H. *Radiocarbon* **5**, 163-202 (1963).
14. Vogel, J. & Visser, E. *Radiocarbon* **23**, 43-80 (1981).

A unique colour and polarization vision system in mantis shrimps

N. Justin Marshall

School of Biological Sciences, University of Sussex, Falmer, Brighton, Sussex BN1 9QG, UK

The apposition compound eyes of mantis shrimps (stomatopods) are divided into three sections, the dorsal and ventral hemispheres and the midband. Many ommatidia of both hemispheres, and all those in the midband, sample the same narrow band in space. The function of the morphologically distinct midband region is not clear, but new evidence suggests that it may be adapted in a unique manner for colour and polarization vision. A series of carotenoid colour filters screen the photopigment and potentially provide a tetrachromatic input for contrast-enhanced vision or true colour vision[1]. The filters are blocks of coloured droplets (red, orange, yellow, purple, pink or blue) within the rhabdoms of two rows of midband ommatidia. The arrangement of tiered microvilli in two other midband rows suggests that they provide a unique form of polarization vision.

In gonodactyloid stomatopods several rows of ommatidia in each hemisphere, and all six rows of midband ommatidia, view the same area in space (Fig. 1). It has been assumed since the work of Exner[2], that upper and lower hemispheres operate together as a rangefinder within each eye, and that this helps to mediate the behaviour for which stomatopods are famous: that is the ballistic smashing or spearing of prey with a pair of raptorial appendages. The midband does not readily fit this scheme and has ommatidia that are very different from those in the hemispheres[3].

Figure 2 summarizes the internal structure of the midband

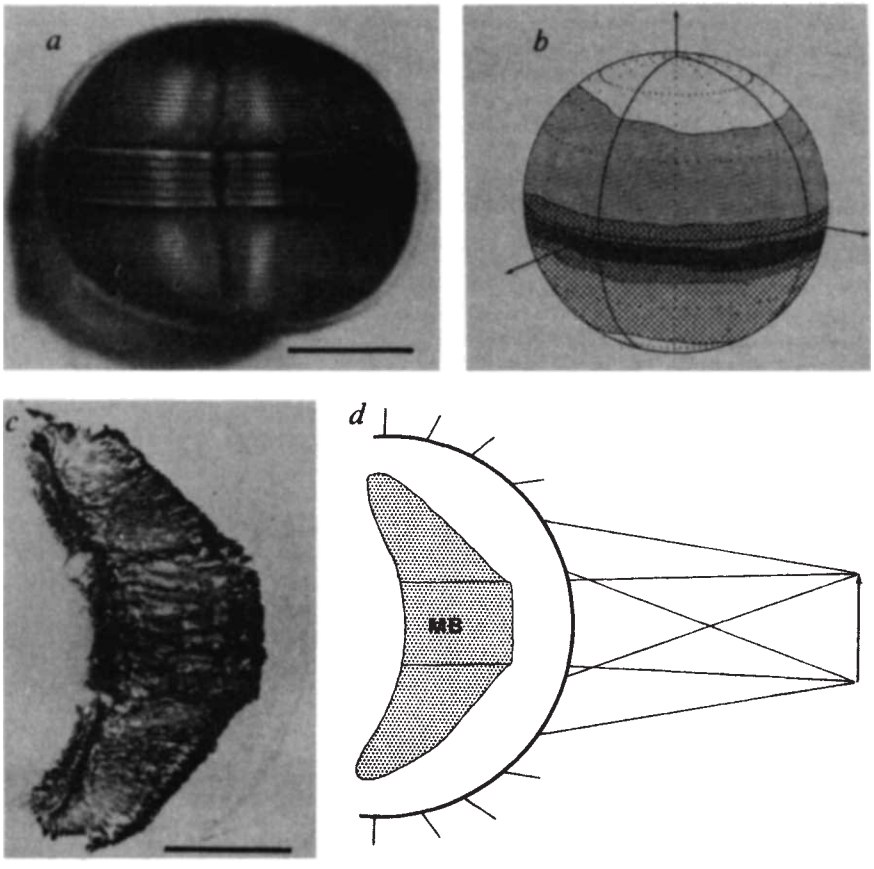
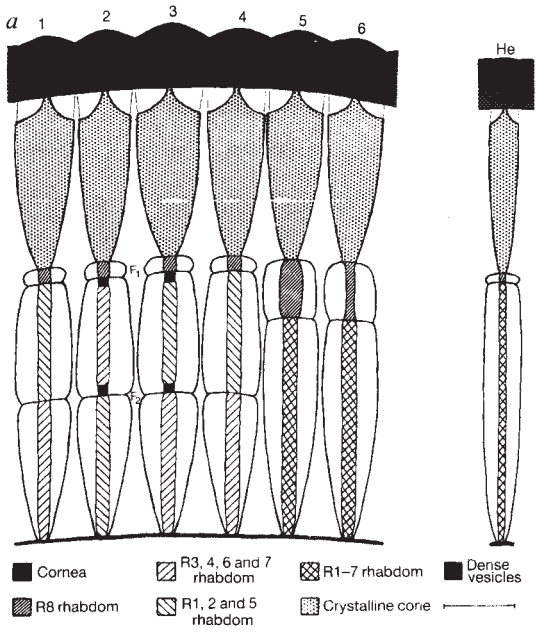


Fig. 1 *a*, The eye of *Odontodactylus scyllarus* showing dorsal and ventral hemispheres and the six ommatidial rows in the midband, typical of gonodactyloid stomatopods. The enlarged facet diameter of the midband suggests a functional difference. The three vertical dark areas (the pseudopupils) are those areas of the eye looking at the observer. Scale, 1.5 mm. *b*, Field of view of the hemispheric and midband regions mapped on to an imaginary globe in space with the eye at the centre and arranged with the midband congruent to the equator. Measurements were made by mapping the pseudopupil position with a goniometer[28]. Small dots, top hemisphere; large dots, bottom hemisphere; solid shading, midband. Note the strip in space where the fields of view of all three areas overlap. In *O. scyllarus*, this strip subtends a maximum angle of 5-15°, but as this method tends to overestimate, the true value could be less. *c*, Sagittal cryosection of eye in *Gonodactylus ternatensis*. Note the amount of space the six midband rows occupy in the retina (the dark crescent) compared to the two hemispheres, each of which contains over 30 ommatidial rows. *d*, Diagram of the visual axes of the column of ommatidia shown in *c*, MB represents the midband region. The visual overlap of the three eye regions is due to ommatidial 'skewing' in each hemisphere towards the midband. The optical intersection of all three eye regions is not to scale.



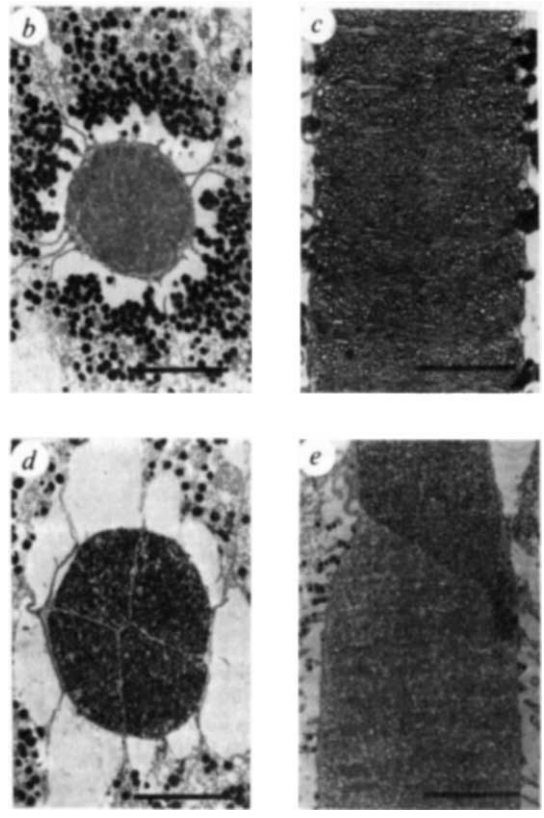


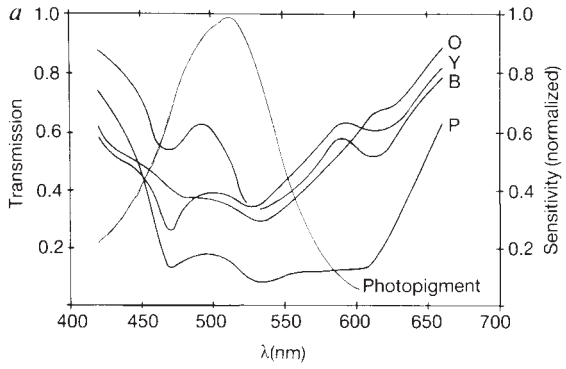
Fig. 2 *a*, Schematic diagram of midband ommatidia in saggital section for a gonodactyloid stomatopod; rows are labelled 1-6 dorsal—ventral (see Fig. 1). The rhabdom width is drawn as three times the actual width, and pigment cells are omitted for clarity. Note the tiered arrangement in all rhabdoms, in particular the two tiers in rows 2 and 3 labelled F1 and F2. The rhabdomeres contributing to each tier are indicated in the key and, apart from R8, follow no previous labelling system. Rows 5 and 6 are the subject of Fig. 4; however the relative length of R8 is shown here. A typical hemispheric ommatidium (He) is included for comparison. Scale, 150 μm. *b*, Transmission electron micrograph (TEM) of the most proximal tier in midband row 3 of *G. chiragra* (transerve section). Note the four cells forming the rhabdom in this tier, R3, 4, 6 and 7. The three cells of the overlying tier, R1 2 and 5, have formed axons. Scale, 6 μm. c, TEM of midband row three, proximal tier in *G. chiragra*; longitudinal section. The layered, orthogonal arrangement of microvilli is typical of most crustaceans and is found in all stomatopod midband and hemispheric rhabdoms, including R8 cells, with the exception of the R8 tier in rows 5 and 6. Scale, 4 μm. *d*, TEM of midband row 2 in *G. chiragra*, transverse section of F2 level; dense-staining vesicles are produced by R3, 4, 6 and 7 in place of rhabdom microvilli Scale , 3 μm. *e*, TEM of midband row 3 in *O. scyllarus*, longitudinal section of F2 and most proximal tier in row three. F1 and 2 tiers in rows 2 and 3 consist of densely staining vesicles that must filter light reaching more proximal layers. Scale, 5 μm.

obtained from ultrathin and semithin serial sections. The rhabdoms of the lower and upper hemispheric regions are typical of many crustaceans and consist of eight retinular cells (R1-8) that are arranged in two tiers, with one cell (R8) producing the upper shorter layer (usually 5-10% of the total rhabdom length) and R1-7 producing the remainder[4]. Midband rhabdoms differ in three respects: (1) they are an average of 10% longer and 100% wider than those of the hemispheres; (2) rows 5 and 6 are two-tiered, like the ommatidia in the hemispheres, but the upper R8 layer is unusually long (20% of the total length) and both tier levels are unlike other rhabdoms in transverse section, being square at the R1-7 level and lozenge-shaped at the R8 level (Fig. 4); (3) there are three tiers in rows 1 and 4, and five in rows 2 and 3. Two of the tier levels in rows 2 and 3 are made up of electron-dense vesicles (Fig. 2), rather than microvilli, and these vesicles contain carotenoid pigments. Thus, in these two rows rhabdom blocks alternate with colour filters.

More than two retinal tiers have not been reported for any other crustacean; but some insects, butterflies and dragonflies for instance, have three-layered retinae in at least a subsection of their eyes[5,6]. A possible function of such an arrangement is to assist in colour vision[7]. All colour vision systems studied to date (in both vertebrates and invertebrates) use multiple photopigments with different spectral sensitivities. Each photopigment is found in discrete receptor populations, and colour opponency is set up by comparison between the receptor types. Where tiered rhabdoms occur, one block of rhabdom is screened by the photopigment of another, and this filter action sharpens the spectral peak of the more proximal rhabdom, potentially improving the contrast between colours[8].

Some vertebrates use another type of filter that has photostable coloured oil droplets positioned over the retinal cones[9,10], so that light must pass through these before reaching the photopigment. The pigeon, for instance has red, yellow, green and clear droplets in its eye distributed in specific areas of its retina. The colouring in these structures comes from carotenoid pigments[11], and with two or more differently coloured droplets acting as filters, it would in principle be possible to achieve colour vision with a single photopigment[10]. All animals with oil droplets so far examined, however, possess several photopigments, and retain colour vision even if the droplets are bleached of colour[12]. Birds may use these filters either to sharpen their existing spectral sensitivities or to divide the retina into areas with which to perform different visual tasks, such as distinguishing foliage or looking through water surfaces[13].

The electron-dense droplet structures in stomatopods are very much larger than vertebrate oil droplets (Fig. 3). In cryosection they appear brightly coloured (red, yellow, orange, pink and, most surprisingly, blue and purple) and, from the three-peaked spectral characteristics, these pigments are likely to be carotenoids (Fig. 3)[14]. They are situated at two specific levels in all rhabdoms of rows 2 and 3 of the midbands: F1 and F2 in Fig. 2. Interestingly, the combination of colours at the F1 and F2 level have so far proved to be species-specific and may relate to body colouration. Most stomatopods are found on coral reefs and, like many tropical reef species, can be brightly coloured; *Odontodactylus scyllarus*, for instance, has striking red, yellow and blue markings on a predominantly green body. *Gonadactylus* sp. are often drab greens and browns, but possess brightly coloured meral spots on the inner surface of their



b

Species	Row 2 F_1	Row 2 F_2	Row 3 F_1	Row 3 F_2
Odontodactylus scyllarus	Yellow 16	Yellow 40	Red 16	Red 56
Gonodactylus ternatensie	Yellow 16	Orange 8	Purple 16	Blue 8
Gonodactylus smithii	Yellow 16	Orange 16	Salmon 16	Blue 8
Gonodactylus chiragre	Yellow 32	Red 8	Clear 10	Blue 8

Fig. 3 *a*, Transmission spectra of the four carotenoid filter blocks in *G. ternatensis*, the absorbance of which is corrected for relative filter length: O, orange; Y, yellow; B, blue and P, purple. Measurements were made using a microscope modified as a microspectrophotometer[29]. Normalized sensitivity spectrum of the photopigment probably present in R1–7 of the midband[19] is represented by the thin line. With such a combination, tetrachromatic vision is possible. *b*, Summary of filter colours at the F1 and F2 level in some species studied to date. Such filters are only found in rows 2 and 3 and their combination seems to be species specific. Also the length of the filters, and therefore the amount of light they let through to successive layers, varies between species. Average lengths in microns of filter blocks are given for 10-cm-long individual of each species.

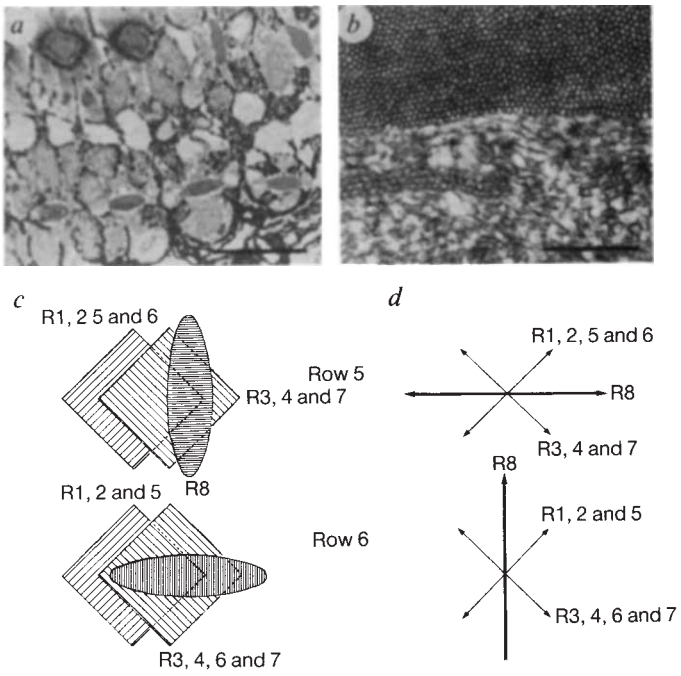


Fig. 4 *a*, Semithin transverse section of midband rows 5 (top) and 6 in *O. scyllarus*: a cross-section taken to show both R8 and R1–7 rhabdom tier levels (R1–7 visible in row 5 only). R8 and R1–7 are both unusual in cross-section, being lozenge and rhombus-shaped respectively. Scale, 40 μm. *b*, TEM of longitudinal section at R8/R1–7 join in *O. scyllarus*. Note the regular closely packed microvilli of R8 running in a single direction only and perpendicular to the lozenge long axis in *a*, which is typical of known polarization vision systems[24]. R1–7 microvilli are orthogonally arranged in that two groups of retinular cells carry microvilli perpendicular to each other. Scale, 1 μm. *c*, Diagram of E-vectors, maximally absorbed in the complete rhabdom, showing single-direction R8 on top of orthogonal R1–7. Microvillus arrangement is shown on the left-hand side, and probable maximal sensitivity on the right. These directions do not change from ommatidium to ommatidium, nor do the individual ommatidia twist. The R8 cells (heavy arrows) may be part of a two-channel system with output comparison between rows 5 and 6. Alternatively, each of rows 5 and 6 may contain a separate three-channel system with output comparison: in row 6 for instance, between R8, R1, 2 and 5 and R3, 4, 6 and 7. Interestingly R8 cells are thought to contain short-wavelength photopigment (around 440 nm), and in many animals polarization sensitivity is greatest at this end of the spectrum[30].

raptorial appendages. These seem to be the only visible means of species identification and animals display these spots in inter- and intraspecific communication[15]. Species identification may be a function of ommatidial rows 2 and 3, and this would imply that the filters are concerned with true colour vision rather than contrast enhancement.

Amongst other arthropods, red pigment granules have been found in the retinae of some pierid butterflies, but most of this pigment is outside the rhabdom, where it could exert only a lateral filtering effect on the photopigment present[16]. In several species of fly, certain R7 rhabdomeres contain a photostable carotenoid pigment distributed in a diffuse manner within the microvillar portion of the cell[17]; here it may protect the R7 microvilli against photooxidation or shift the maximal spectral sensitivity of the underlying R8 cell[18]. The carotenoid pigment of stomatopods forms dense plugs situated in the rhabdom and must therefore act as a filter, altering the spectral quality of light reaching more proximal layers. As the length of the filters varies within and between species (Fig. 3), the strength of this effect will vary.

It is not clear at present whether or not stomatopods use several photopigments, like vertebrates. It is quite possible that they may achieve colour vision with a single photopigment. In most crustaceans, including stomatopods, one photopigment peaking at around 510 nm has been found in R1–7 (ref. 19). Recently Cummins and Goldsmith[20] demonstrated a violet-sensitive 440-nm photopigment in crayfish that is confined to the short distal R8 cells, and this may be the case for R8 cells in other crustacean species[21]. Where the R8 cell is short, it will probably have little effect on the spectral sensitivity of the main R1–7 rhabdom tier. Stomatopods possessing two or more differently coloured carotenoid filters could construct a colour vision system based on a 510-nm photopigment in R1–7. Microspectrophotometry of the carotenoid filters (Fig. 3) indicates that they act as highly specific cutoff filters, restricting the spectral range of light incident on more proximal layers to long and short wavelengths (a rough indication of the resulting sensitivities would be the product of the transmission curves and sensitivity curve in Fig. 3). Rhabdomeres in midband rows 1 and 4, which have no carotenoid filters, might be compared against rows 2 and 3 to provide colour opponency at the neuronal level.

The structure of the remaining midband rows 5 and 6 strongly suggests that they operate as polarization analysers (Fig. 4). In many arthropods, polarization vision is thought to be based on rhabdomere structure, with most rhabdoms being maximally sensitive to light whose E-vector is parallel to the microvilli they carry[22]. Crustacean microvilli are typically arranged

orthogonally in alternating layers[3] (Fig. 2*c*), and in these a two-channel system could be envisaged, with the two rhabdomere subsets being most sensitive to light with **E**-vectors at right angles to each other. Labhart has recently demonstrated such a system in the dorsal rim area of crickets where rhabdom microvilli are also arranged orthogonally[23].

R8 in rows 5 and 6 have microvilli arranged in one direction only and at right angles to each other. This suggests a two-channel system involving signal comparison between R8 cells in rows 5 and 6 (Fig. 4). Interestingly, R8 cells in these two rows are longer than in the rest of the eye and possess closely packed highly regular microvilli of the sort used in some known polarization vision systems[24]. They may therefore have a substantial polarizing effect on light reaching R1-7. It is possible that gonodactyloid stomatopods actually use two different three-channel systems. The rhabdom tier beneath R8, R1-7, in both rows 5 and 6, contains orthogonal microvilli whose two directions of maximal sensitivity are at 45° to those of R8. Comparison of three channels (R8; R1, 2 and 5 and R3, 4, 6 and 7 in row 6 ommatidia + R8; R1, 2, 5 and 6 and R3, 4 and 7 in row 5, see Fig. 4) eliminates the possible points of ambiguity found in a two-channel system[25]. This would allow a more accurate appraisal of polarized light in the environment. In stomatopods polarization vision might be involved in the analysis of shiny surfaces or particle-scattered light[1]. In both cases this may help increase the contrast with the background of objects being examined, such as potential predators or prey.

It seems likely that some stomatopods possess a potentially complex colour and polarization vision system in the midband region of their eyes. Before this can be stated with any certainty however, some behavioural evidence is needed to show positive discrimination between colours and between **E**-vectors. Such colour or polarization vision systems involving sequential filtering reduce the light reaching the rhabdom and the large lenses, and rhabdoms of the midband evidently compensate for this by increasing the light-gathering capabilities of the ommatidia. The narrow 'trinocular' strip viewed by each eye is moved through space by complex, largely independent eye movements[26] so that it may be swept over an object which the midband analyses for colour and polarization properties, and the two hemispheres for distance. Like many arthropods[27], stomatopods have come up with an ingenious series of compromises and modifications to different areas of their eyes to perform different tasks.

Many thanks to Professor M. F. Land for invaluable tuition, guidance and criticism. Also thanks to Dr T. S. Collett and Dr R. C. Hardie for useful discussion and criticism, and to Dr J. R. Thorpe for technical advice. This study was supported by a grant from the SERC (UK).

Received 14 March; accepted 29 April 1988.

1. Lythgoe, J. N. *The Ecology of Vision* (Clarendon, Oxford, 1979).
2. Exner, S. in *Die Physiologie der Facettirten Augen von Krebsen und Insecten* (Deuticke, Leipzig and Wien, 1891).
3. Schiff, H. *Biol. Bull.* **170**, 461-480 (1986).
4. Waterman, T. H. in *Identified Neurons and Behaviour of Arthropods* (ed. Hoyle, G.) 317-386 (Plenum, New York, 1977).
5. Laughlin, S. B. & McGinness, S. *Cell Tiss. Res.* **188**, 427-447 (1978).
6. Schlecht, P. *et al. J. comp. Physiol.* **123**, 239-243 (1978).
7. Meinertzhagen, I. A. *et al. J. comp. Physiol.* **151**, 295-310 (1983).
8. Bowmaker, J. K. *Trends neuro. Sci.* **6**, 41-43 (1983).
9. Jane, S. D. & Bowmaker, J. K. *J. comp. Physiol.* **A162**, 225-235 (1988).
10. Bowmaker, J. K. *Trends neuro. Sci.* **3**, 196-199 (1980).
11. Liebman, P. A. & Granada, A. M. *Nature* **253**, 370-372 (1975).
12. Wallman, J. in *Neuronal Mechanisms of Behaviour in the Pigeon* (eds Granada, A. M. & Maxwell, J. H.) 327-351 (Plenum, New York, 1979).
13. Munz, W. R. A. in *Handbook of Sensory Physiology* Vol. VII/1 (ed. Autrum, H.) 529-565 (Springer, Berlin, 1972).
14. Fox, H. M. & Vevers, G. in *The Nature of Animal Colours* 62-80 (Sidgwick & Jackson, London, 1966).
15. Caldwell, R. L. & Dingle, H. *Scient. Am.* **234**, 80-89 (1976).
16. Ribi, W. A. *J. comp. Physiol.* **132**, 1-9 (1979).
17. Kirschfeld, K. *et al. J. comp. Physiol.* **125**, 275-284 (1978).
18. Kirschfeld, K. *et al. J. comp. Physiol.* **162**, 421-433 (1988).
19. Cronin, T. W. *J. comp. Physiol.* **156**, 679-687 (1985).
20. Cummins, D. R. & Goldsmith, T. H. *J. comp. Physiol.* **142**, 199-202 (1981).
21. Martin, F. G. & Mote, M. I. *J. comp. Physiol.* **145**, 549-554 (1982).
22. Snyder, A. W. & Laughlin, S. B. *J. comp. Physiol.* **100**, 101-166 (1975).
23. Labhart, T. *Nature* **331**, 435-437 (1988).
24. Saibil, H. & Hewat, E. *J. cell. Biol.* **105**, 19-28 (1987).
25. Bernard, G. D. & Wehner, R. *Vision Res.* **17**, 1019-1028 (1977).
26. Cronin, T. W. *J. expl. Biol.* (in the press).
27. Land, M. F. in *Facets of Vision* (ed. Hardie, R. C. & Stavenga, D. G.) (Springer, Berlin, in the press).
28. Horridge, G. A. *Phil. Trans. R. Soc.* **285**, 1-59 (1978).
29. Knowles, A. & Dartnall, H. J. in *The Eye* Vol. 2B (ed. Davson, H.) (Academic, London and New York, 1977).
30. Wehner, R. *J. comp. Physiol.* **161**, 511-531 (1987).

Platelet-derived growth factor promotes division and motility and inhibits premature differentiation of the oligodendrocyte/type-2 astrocyte progenitor cell

Mark Noble*§, Kerren Murray†, Paul Stroobant*, Michael D. Waterfield* & Peter Riddle‡

* Ludwig Institute for Cancer Research, 91 Riding House Street, London W1P 8BT, UK
† Institut Merieux, 69260 Charbonniers les Bains, Lyon, France
‡ Imperial Cancer Research Fund, Lincolns Inn Fields, London WC2A 3PX, UK

The mitogens which modulate cell–cell interactions during development of the central nervous system are unknown. One of the few interactions sufficiently well understood to allow identification of such molecules involves the two glial lineages which make up the rat optic nerve. One population of glial cells in this tissue, the type-1 astrocytes[1], secrete a soluble factor(s) which promotes division of a second population of bipotential oligodendrocyte/type-2 astrocyte (O-2A) progenitor cells[2]; these progenitors give rise to oligodendrocytes, which myelinate large axons in the CNS, and type-2 astrocytes, which enwrap bare axons at nodes of Ranvier[3,4]. Type-1 astrocytes also promote progenitor motility[5], and inhibit the premature differentiation of progenitors into oligodendrocytes which occur when these cells are grown in the absence of type-1 astrocytes[2,3]. We have now found that platelet-derived growth factor mimics the effects of type-1 astrocytes on O-2A progenitor cells, and antibodies to PDGF block the effects of type-1 astrocytes.

To analyse the molecular basis for the effects of type-1 astrocytes, we first looked at the ability of diverse growth factors (epidermal growth factor, basic fibroblast growth factor, transforming growth factor-α, transforming growth factor-β, glial growth factor, platelet-derived growth factor and interleukins 1, 2 and 3) to modulate DNA synthesis and differentiation of O-2A progenitor cells in cultures derived from optic nerves of 7-day-old rats. Such cultures contain many O-2A progenitors, but few type-1 astrocytes, reducing the possibility that any effects might be indirectly mediated by type-1 astrocytes[2,3]. Individual cell-types were identified as described (refs 1-5; legend to Table 1). Only platelet-derived growth factor (PDGF) was found to mimic the effects of type-1 astrocytes.

PDGF promoted DNA synthesis in O-2A progenitor cells as effectively as type-1 astrocytes or astrocyte-conditioned medium (ACM) (Table 1). In the presence of 10 ng ml^{-1} PDGF, 70-80% of the O-2A progenitors in cultures of 7-day-old optic nerve were labelled with [^{3}H] thymidine during a 20 h pulse. With the best PDGF preparations, the response was maximal at $\leq$5 pM PDGF (Fig. 1*a*). As all batches of PDGF examined gave maximal responses at $\leq$10 ng ml^{-1}, we used this concentration in most experiments.

PDGF also inhibited the rapid differentiation of progenitor cells into oligodendrocytes that is otherwise seen when these

§ To whom correspondence should be addressed.

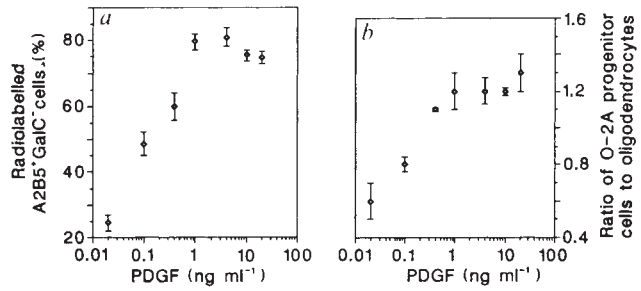


Fig. 1 Nanogram quantities of PDGF induce DNA synthesis and inhibit premature differentiation in O-2A progenitors. Cells from optic nerves of 7-day-old rats were grown in the presence of varying concentrations of PDGF for 48 h (see legend to Table 1 for culture details). *a*, Proportion of O-2A progenitors in which ≥10 silver grains were counted above the cell nucleus following immunostaining and autoradiography (see also Table 1). *b*, Ratio of A2B5$^+$, GalC$^-$ O-2A progenitor cells to GalC$^+$ oligodendrocytes (see also Table 2). The effects of PDGF reached a maximum at 1 ng ml^{-1}, and the effects of PDGF at 0.02 ng ml^{-1} were generally indistinguishable from control values (data not shown), although in some experiments no radiolabelled A2B5$^+$ cells were found in DMEM-BS. Note that although 25% of O-2A progenitor-like cells had radiolabelled nuclei even with 0.02 ng ml^{-1} PDGF, the absolute numbers of such cells were very low, as most O-2A lineage cells differentiated into oligodendrocytes in these conditions. At these low concentrations, as in the absence of PDGF, almost all O-2A progenitors differentiated into oligodendrocytes within an additional 24 h of growth *in vitro* (unpublished observations). Figures are averages of triplicate coverslips ±s.d. for the most effective batch of PDGF tested.

Table 1 PDGF induces DNA synthesis in O-2A progenitor cells

Condition	O-2A progenitor cells with radiolabelled nuclei (%) 24 h	48 h
DMEM-BS	7.6 ± 1.8	1.4 ± 0.7
Astrocyte monolayers	65.8 ± 5.4	72.3 ± 5.1
ACM	71.2 ± 4.3	70.6 ± 4.7
PDGF, 10 ng ml^{-1}	83.6 ± 3.1	79.8 ± 5.2

The PDGF used in these experiments was the porcine B-chain homodimer[12], but all results were confirmed using a preparation of human A-B heterodimer (gift of Thomas Deuel). The percentages are averages of triplicate coverslips ±s.d.; similar results were obtained in six separate experiments. Optic-nerve cells derived from 7-day-old rats were plated either in a 25 μl drop on poly-L-lysine (PLL-coated glass coverslips (13 mm) at 10,000 cells per coverslip in Dulbecco's modified Eagle's medium plus transferrin, insulin, selenium, progesterone, thyroxine, tri-iodothyronine, putrescine and bovine serum albumin (DMEM-BS), as described[2,3], or directly onto monolayers of irradiated type-1 astrocytes growing in DMEM-BS; dissociated optic-nerve cells and monolayers of purified type-1 astrocytes were prepared as previously[3]. Cells were allowed to adhere for 20 min before being fed to 0.5 ml with DMEM-BS, a 1:1 mixture of DMEM-BS plus DMEM-BS which had been conditioned for 48 h by astrocyte monolayers, or PDGF (10 ng ml^{-1}). [^{3}H]-thymidine was either added immediately for a 24 h pulse, or after a delay of 24 h; in the latter case cultures received fresh aliquots of mitogen with the [^{3}H]-thymidine addition. After radiolabelling, cultures were stained with A2B5 (ref. 13) and antigalactocerebroside (GalC, ref. 14) monoclonal antibodies, followed by labelling with appropriate conjugates, fixation and processing for autoradiography[3]. O-2A progenitors were defined as A2B5$^+$ GalC$^-$ cells, oligodendrocytes as GalC$^+$ cells, and type-2 astrocytes as cells labelled with A2B5 antibody and with rabbit antiserum against the astrocyte-specific glial fibrillary acidic protein (GFAP, refs 15, 16). Staining on separate coverslips confirmed that >98% of the O-2A lineage cells in the above experiments were GFAP$^-$.

Table 2 PDGF inhibits the premature differentiation of O-2A progenitor cells into oligodendrocytes *in vitro*

Condition	O-2A lineage cells with a progenitor phenotype (%) A2B5$^+$, GalC$^-$ antigenic phenotype	Bipolar morphology
DMEM-BS	2.5 ± 1.0	0
ACM	52.8 ± 4.6	46.4 ± 3.1
PDGF, 10 ng ml^{-1}	56.3 ± 2.6	43.2 ± 0.9

Almost all O-2A progenitors differentiated into GalC$^+$ oligodendrocytes within 72 h when grown in the absence of mitogen. In contrast, cultures grown in the presence of ACM or PDGF contained many cells which did not differentiate into GalC$^+$ oligodendrocytes during this same period. Staining of parallel cultures confirmed that >98% of the A2B5$^+$, GalC$^-$ cells were also GFAP$^-$. The percentages are ±s.d. for triplicate coverslips. Optic-nerve cells were grown as in Table 1 in the conditions indicated. Cultures were labelled after 72 h with A2B5 and anti-GalC antibodies and appropriate conjugates and examined to determine (1) the proportion of total O-2A lineage cells (that is, all cells labelling with either or both antibodies[2-4]) which were GalC$^-$, and (2) the proportion of these cells which expressed the bipolar morphology characteristic of O-2A progenitor cells at their earliest stages of development (ref. 5, I. Sommer and M.N. manuscript in preparation). As described in the text and previously[5], these bipolar cells are migratory and are also the major source of dividing cells in these cultures.

cells are grown in the absence of type-1 astrocytes. After 72 h of growth in chemically defined medium, almost all O-2A lineage cells differentiated into oligodendrocytes (refs 2, 3; Table 2). In contrast, after 48 or 72 h of growth in the presence of PDGF or ACM, more than half of the O-2A lineage cells retained the antigenic phenotype of O-2A progenitors (Table 2, Fig. 1*b*), and >40% of these antigenically defined cells expressed the bipolar morphology characteristic of dividing and motile O-2A progenitor cells (Table 2; see also below and ref. 5). The dose-response curve for inhibition of progenitor differentiation was identical to that for the induction of DNA synthesis (Fig. 1*a* and 1*b*). As discussed below, however, this inhibition did not absolutely preclude the differentiation of dividing progenitors into oligodendrocytes.

Cell division, differentiation and motility were examined continuously for several days using time-lapse microcinematography. Cultures grown in the presence of PDGF, like those grown in ACM[5], contained many motile dividing cells with the characteristic bipolar morphology of O-2A progenitors; in parallel cultures, all cells with this morphology expressed the antigenic phenotype of O-2A progenitors (that is, A2B5$^+$, galactocerebroside (GalC)$^-$, glial fibrillary acidic protein (GFAP)$^-$. Progenitors grown in the presence of PDGF migrated with an average speed of 24.6 ± 5.4 μm h^{-1} (40 cells examined), comparable with 21.4 μm h^{-1} for those grown in ACM[5] and contrasting with the lack of migration of cells grown in chemically-defined medium. Cells grown in the presence of PDGF or ACM also had similar cell-cycle lengths (19.7 ± 6.4 h in PDGF, 17.7 ± 4 h in ACM; n = 26 and 31 cells analysed, respectively).

We often observed differentiation of motile bipolar O-2A progenitor cells into multipolar cells with the characteristic morphology and lack of motility of oligodendrocytes in cultures grown in the presence of PDGF and followed by time-lapse microcinematography. Immunolabelling of parallel cultures demonstrated that such cells were indeed oligodendrocytes expressing the antigenic marker GalC. Cultures grown for 72 h in chemically-defined medium containing PDGF contained few type-2 astrocytes (antigenically A2B5$^+$, GFAP$^+$) unless fetal calf serum (FCS) was also added to the cultures. When grown in medium containing 10% FCS for 72 h, however, 90% of all A2B5$^+$ cells were GFAP$^+$ type-2 astrocytes regardless of whether

Table 3 Anti-PDGF antibodies block response of O-2A progenitors to type-1 astrocytes and PDGF, but not to FGF

	Radiolabelled O-2A progenitors (%)	
	No anti-PDGF	Plus anti-PDGF
ACM	63+4.2%	<1%
PDGF, 5 ng/ml	69+3.7%	<1%
FGF, 5 ng/ml	40+4.2%	44.5+2.5%

Cells were grown as in Table 1. Cultures received either the indicated mitogen, or mitogen plus 50 μg of rabbit anti-PDGF antiserum (purified Ig fraction, a kind gift of C. Heldin). When antibody was added, the ACM or non-conditioned medium with mitogen was preincubated with antibody for at least 1 h before addition to the cells. Cells were labelled with [^{3}H]-thymidine, immunolabelled, processed for autoradiography and scored as in Table 2.

the medium contained PDGF (data not shown). Thus, PDGF neither inhibited nor induced differentiation of O-2A progenitor cells into type-2 astrocytes, nor did it preclude differentiation of O-2A progenitor cells into oligodendrocytes. In these respects also, the effects of PDGF were identical to the effects of type-1 astrocytes.

To determine whether the effects of type-1 astrocytes were mediated by PDGF, we treated ACM, or medium containing PDGF or fibroblast growth factor (FGF) with affinity-purified anti-PDGF antibody. (As will be discussed elsewhere, FGF is also a mitogen for O-2A lineage cells, but does not mimic the effects of type-1 astrocytes on motility and differentiation.) These antibodies blocked the effects of ACM and PDGF, but did not block FGF-induced DNA synthesis in O-2A progenitors (Table 3), indicating that blocking was not due to toxic effects of the antibody.

Type-1 astrocytes also support the appropriately timed differentiation of embryonic O-2A progenitors grown *in vitro*[6]. As discussed in the accompanying paper[7], astrocyte-derived PDGF seems to play a key role in this effect and is by itself sufficient to promote the synchronous differentiation of clonally related and dividing progenitor cell families. Thus, in these respects also, PDGF completely mimics the effects of type-1 astrocytes.

The ability of a single molecule to replace type-1 astrocytes in modulating O-2A progenitor development *in vitro* suggests that these cells have a complex and constitutive behavioural phenotype, controlled by processes internal to the progenitors themselves. Progenitors stimulated to divide by an appropriate mitogen appear to be intrinsically migratory cells, with a bipolar morphology, which cease migration upon differentiating into multipolar oligodendrocytes; this differentiation seems to be controlled, at least in part, by internal clocks which may function by counting cell divisions. This programme does not, however, include astrocyte differentiation, which requires a separate inducing factor[10].

The observations that PDGF induced DNA synthesis at picomolar concentrations, the identical effects of PDGF and type-1 astrocytes on the division, differentiation and motility of O-2A progenitors and the ability of anti-PDGF antibodies to block the activity of astrocyte-conditioned medium all indicate that the astrocyte activity is a PDGF-like substance. In addition, messenger RNA for PDGF has been identified in purified astrocytes and these cells have been found to secrete PDGF *in vitro*[17] (W. Richardson, N. Pringle, M. Moseley, B. Westermark, & M. Dubois-Dalcq, manuscript submitted). Together, our studies indicate that PDGF may play an important role in gliogenesis in the CNS.

The ability of PDGF to promote division and migration of O-2A progenitor cells may be of particular interest in light of observations that PDGF can act as a chemotactic agent[11]. During embryogenesis, O-2A progenitors appear to populate the optic nerve by migrating from a germinal zone in or near the optic chiasm along the nerve towards the eye[5]. This directional migration could be due to movement along a gradient of a chemotactic substance, such as PDGF. In the adult animal, the ability of cells in a lesion site to secrete compounds which promote O-2A progenitor migration and division could be of value in the repair of demyelinating damage in the CNS. The controlled application of PDGF, or other chemotactic mitogens, might enhance these repair processes.

Received 25 February; accepted 29 April 1988.

1. Raff, M. C., Abney, E. R., Cohen, J., Lindsay, R. & Noble, M. *J. Neurosci.* **3**, 1289-1300 (1983).
2. Noble, M. & Murray, K. *EMBO J.* **3**, 2243-2247 (1984).
3. Raff, M. C., Miller, R. H. & Noble, M. *Nature* **303**, 390-396 (1983).
4. ffrench-Constant, C. & Raff, M. C. *Nature* **323**, 335-338 (1986).
5. Small, R. K., Riddle, P. & Noble, M. *Nature* **328**, 155-157 (1987).
6. Raff, M. C., Abney, E. R. & Fok-Seang, J. *Cell* **42**, 61-69 (1985).
7. Raff, M. C., Lillien, L. E., Richardson, W. D., Burne, J. F. & Noble, M. D. *Nature* **333**, 562-565.
8. Temple, S. & Raff, M. C. *Cell* **44**, 773-779 (1986).
9. Mirsky, R. *et al. J. Cell Biol.* **84**, 483-494 (1980).
10. Hughes, S. & Raff, M. C. *Development* **101**, 157-167 (1987).
11. Ross, R., Raines, E. W. & Bowen-Pope, D. F. *Cell* **46**, 155-169 (1986).
12. Stroobant, P. & Waterfield, M. D. *EMBO J.* **3**, 2963-2967 (1984).
13. Eisenbarth, G. S., Walsh, F. S. & Nirenberg, M. *Proc. natn Acad. Sci. U.S.A.* **76**, 4913-4917 (1979).
14. Ranscht, B., Clapshaw, P. A., Price, J., Noble, M. & Seifert *Proc. Natn. Acad. Sci. U.S.A.* **79**, 2709-2713 (1982).
15. Bignami, A., Eng, L. F., Dahl, D. & Uyeda, C. T. *Brain Res.* **43**, 429-435 (1972).
16. Pruss, R. M. *Nature* **280**, 688-690 (1979).
17. Richardson, W. D., Pringle, N., Mosley, M. J., Westermark, B. & Dubois-Dalcq, M. *Cell* **53**, 309-319 (1988).

Platelet-derived growth factor from astrocytes drives the clock that times oligodendrocyte development in culture

Martin C. Raff*†, Laura E. Lillien*†, William D. Richardson†, Julia F. Burne*† & Mark D. Noble‡

* MRC Developmental Neurobiology Programme and † Biology Department, Medawar Building, University College London, London WC1E 6BT, UK
‡ Ludwig Institute for Cancer Research, 91 Riding House Street, London W1P 8BT, UK

The various cell types in a multicellular animal differentiate on a predictable schedule but the mechanisms responsible for timing cell differentiation are largely unknown. We have studied a population of bipotential glial (O-2A) progenitor cells in the developing rat optic nerve[1] that gives rise to oligodendrocytes beginning at birth and to type-2 astrocytes[2] beginning in the second postnatal week[3]. Whereas, *in vivo*, these O-2A progenitor cells proliferate and give rise to postimitotic oligodendrocytes over several weeks[4,5], in serum-free (or low-serum) culture they stop dividing prematurely and differentiate into oligodendrocytes within two or three days[1,6,7]. The normal timing of oligodendrocyte development can be restored if embryonic optic-nerve cells are cultured in medium conditioned by type-1 astrocytes[8], the first glial cells to differentiate in the nerve[3]: in this case the progenitor cells continue to proliferate, the first oligodendrocytes appear on the equivalent of the day of birth, and new oligodendrocytes continue to develop over several weeks, just as *in vivo*[7]. Here we show that platelet-derived growth factor (PDGF) can replace type-1-astrocyte-conditioned medium in restoring the normal timing of oligodendrocyte differentiation *in vitro* and that anti-PDGF antibodies inhibit this property of the appropriately conditioned medium. We also show that PDGF is present in the developing optic nerve. These findings suggest that type-1-astrocyte-derived PDGF drives the clock that times oligodendrocyte development.

Table 1 Effect of PDGF on O-2A progenitor-cell proliferation and differentiation into oligodendrocytes *in vitro*

Age of optic nerve	PDGF (ng ml^{-1})	Days in culture	Number of O-2A progenitor cells	Number of oligo-dendrocytes
E17	0	2	8±5	15±7
E17	0	3	5±3	19±8
E17	2	2	51±8	0
E17	2	3	117±21	0
E17	2	4	214±38	4±2
E18	0	2	8±3	55±7
E18	2	2	135±17	0
E18	2	3	317±45	10±4
E19	0	1	14±2	70±4
E19	2	1	120±9	0
E19	2	2	199±28	2±1
E18	2	2	148±13	0
E18	2	3	287±23	12±6
E18	10	2	133±11	0
E18	10	3	278±37	8±4

Optic nerves from embryonic S/D rats were dissociated into single cells and cultured on poly-D-lysine (PDL)-coated glas coverslips (about 30,000 cells per coverslip) in Dulbecco's modified Eagle's medium (DMEM) supplemented with glucose, insulin, transferrin, bovine serum albumin, progesterone, putrescine, thyroxine, tri-idodothyronine and 0.5% FCS as described[7]. Purified human PDGF (R and D Systems, Inc.) was added at the start of the culture. After 1-4 days, the cells were fixed in 4% paraformaldehyde in phosphate-buffered saline (*p*H 7.5) for 5 min at room temperature and stained successively with monoclonal anti-galactocerebroside (GC) antibody[19] (ascites fluid diluted 1:1000), fluorescein-coupled goat anti-mouse IgG3 (G anti-IgG3-Fl, Nordic, diluted 1:100), A2B5 monoclonal antibody[20] (ascites fluid diluted 1:100) and finally rhodamine-coupled goat anti-mouse Ig (G anti-MIg-Rd, Cappel, diluted 1:100); the cells were then post-fixed in acid-alcohol, mounted in glycerol and examined in a Zeiss Universal fluorescence microscope, as described[7]. O-2A progenitor cells were identified by their antigenic phenotype (A2B5$^+$, GC$^-$)[1,15] and characteristic process-bearing morphology[15], while oligodendrocytes were identified as GC$^+$ process-bearing cells[11]. The total numbers of these cells were counted on each coverslip and the results are expressed as means ± s.d. of at least three experiments. The concentration of PDGF required for half-maximal stimulation of O-2A progenitor-cell proliferation was ~0.5 ng ml^{-1} (ref. 9).

We were led to study the effect of PDGF on the timing of oligodendrocyte development *in vitro* by recent evidence that PDGF is a potent mitogen for O-2A progenitor cells in culture[9,10] and that cultured type-1 astrocytes stimulate O-2A progenitor-cell proliferation[8] by secreting PDGF[9]. In the present study, optic nerve cells from embryonic day 17 (E17) Sprague-Dawley (S/D) rats were cultured in medium containing 0.5% fetal calf serum (FCS). As previously reported[7], within two days most of the O-2A progenitor cells in such cultures stopped dividing and differentiated into oligodendrocytes, which were identified by the binding of antibody against galactocerebroside (GC)[11] (Table 1). When human PDGF (R and D Systems, Inc., Minneapolis) was added, however, the O-2A progenitor cells continued to proliferate, doubling in number approximately every day, and the first oligodendrocytes developed after four days, equivalent to the time of birth (Table 1). The same result was obtained if the cells were cultured in type-1-astrocyte-conditioned medium (ACM), as reported previously[7] (data not shown). When E18 or E19 optic nerve cells were cultured in PDGF, the first oligodendrocytes developed after three and two days, respectively, again equivalent to the day of birth (Table 1). The same results were obtained with a fivefold higher concentration of PDGF (Table 1), with human PDGF obtained from Raines and Ross[12], and with porcine PDGF obtained either from R and D Systems or from Stroobant and Waterfield[13] (data not shown).

These results indicate that PDGF can mimic the effects of ACM in restoring in culture the normal timing of oligodendrocyte development observed *in vivo*, raising the possibility that PDGF is the factor in ACM responsible for this activity. To test this possibility, E17 optic nerve cells were cultured in ACM together with an IgG fraction of a goat anti-PDGF antiserum. As shown in Table 2, these antibodies completely blocked the ability of ACM both to stimulate O-2A progenitor-cell proliferation (as reported previously[9]) and to restore the normal *in vivo* timing of oligodendrocyte development in culture (Table 2). The same result was obtained using an IgG fraction of a rabbit antiserum[14] obtained from C.-H. Heldin (data not shown). IgG fractions of goat and rabbit antisera against mouse immunoglobulin, used at the same or tenfold higher concentration, had no such effect (data not shown). The addition of exogenous PDGF together with the anti-PDGF antibodies completely overcame the inhibitory activity of the antibodies (Table 2).

Table 2 Effect of anti-PDGF antibodies on the ability of conditioned medium (ACM or ONCM) to restore the *in vivo* timing of oligodendrocyte development in cultures of E17 optic nerve cells

Conditioned medium	Anti-PDGF antibodies (90 μg ml^{-1})	PDGF (ng ml^{-1})	Number of O-2A progenitor cells	Number of oligo-dendrocytes
none	−	0	12±5	25±11
none	−	2	71±9	0
ACM	−	0	75±13	0
ACM	+	0	15±7	24±8
ACM	+	15	82±18	0
ONCM	−	0	64±8	0
ONCM	+	0	8±2	31±5

Cells from E17 optic nerves were prepared, cultured for 2 days and stained as in Table 1. Purified type-1 astrocytes from newborn rat cerebral cortex were prepared as described[21]; after growing for several weeks in DMEM supplemented with 10% FCS, the astrocytes were grown in the defined medium (with 0.5% FCS) described in Table 1 for 2 days and the medium was collected as ACM. Newborn optic nerve cells (5×10^5 in 2 ml in PDL-coated 3.5 mm Nunc tissue-culture dishes) were cultured as described[1]; after 1 day in DMEM containing 10% FCS, the cultures were switched to defined medium containing 0.5% FCS, which was collected after 1 day as ONCM. The conditioned media were tested on E17 optic-nerve cultures to find the highest dilution that would still restore the normal timing of oligodendrocyte differentiation, and this concentration (which varied from undiluted to diluted 1:10) was used in these experiments. The medium was changed after 24 h, and fresh conditioned medium, anti-PDGF antibody, and human PDGF (R and D Systems, Inc.) were added. The goat anti-human PDGF antibodies (an IgG fraction prepared by ion-exchange chromatography) were purchased from Collaborative Research Inc. (Bedford, Massachusetts); 50 μg ml^{-1} of the antibody completely neutralized the mitogenic activity of 5 ng ml^{-1} of human PDGF for O-2A progenitor cells (data not shown), and for NIH 3T3 cells (according to the Collaborative Research specification sheet). The results are expressed as means ± s.d. of three separate experiments, except for the results with ONCM here they are triplicates of a single experiment.

The ACM used in the present and previous studies[7-9] was derived from cultures of type-1 astrocytes purified from neonatal-rat cerebral cortex. To determine whether type-1 astrocytes in optic nerve cell cultures also secrete PDGF, we cultured E17 optic nerve cells in medium conditioned over a high density culture of newborn optic nerve cells; almost 60% of newborn optic nerve cells are type-1 astrocytes[3]. As shown in Table 2, such optic nerve conditioned medium (ONCM) kept O-2A progenitor cells dividing and prevented these cells from prematurely differentiating into oligodendrocytes; this activity was

completely inhibited by anti-PDGF antibodies. We have previously provided evidence, however, that the endogenous type-1 astrocytes in E17 optic nerve cultures are too few in number and/or are unable to recover quickly enough from the dissociation procedure to produce sufficient mitogen to keep the progenitor cells dividing and to prevent their premature differentiation[7].

To determine whether PDGF is made in the developing optic nerve, we tested an extract of three-week-old rat optic nerve for its ability to stimulate O-2A progenitor cells in culture to incorporate bromodeoxyuridine (BrdU) into DNA before and after the extract was treated with anti-PDGF antibodies. As shown in Table 3, the extract stimulated progenitor cells to incorporate BrdU, and more than 70% of this mitogenic activity was removed by anti-PDGF antibodies. While this suggests that the major mitogen for O-2A progenitor cells in the optic nerve at this age is PDGF, the finding that not all of the mitogenic activity in the extract was removed by the antibodies, or neutralized when progenitor cells were cultured with the extract in the presence of an excess of anti-PDGF antibody (data not shown), suggests that another mitogen(s) is also present.

Table 3 Stimulation of BrdU incorporation in O-2A progenitor cells by optic nerve extract in culture: the effect of anti-PDGF antibodies

Additives	% O-2A progenitor cells labelled with BrdU
None	3±1
PDGF (1ng ml^{-1})	64±4
Optic nerve extract	50±2
Optic nerve extract treated with anti-PDGF antibodies	14±6

Cells from newborn optic nerves were cultured as in Table 1, except that they were maintained without FCS at 5,000 cells per culture. Bromodeoxyuridine (BrdU, 10 μM; Boehringer) was added after 11.5 h, and the cultures were fixed after 48 h and stained with A2B5 antibody, followed by G anti-MIg-F1 (Cappel, 1:100); they were then treated successively with 2 N HCl (to denature the nuclear DNA[22]), 0.1 M $Na_2B_4O_7$, *p*H 8.5 (each for 10 min at room temperature), monoclonal anti-BrdU antibody[23] (culture supernatant diluted 1:5) and finally with G ani-MIgRd. Optic-nerve extract was prepared from three week-old rats as described[18] and used at 220 μg total protein ml^{-1}. For treatment with anti-PDGF antibodies, the extract (264 μg total protein) was incubated with the IgG fraction of goat anti-human PDGF antiserum (135 μg in 0.7 ml of DMEM; Collaborative Research) for 4 h at 4 °C with continuous rotation; protein A-Sepharose (50 μl of swollen gel, Pharmacia) was added and the mixture was incubated for a further 12 h at 4 °C with continuous rotation and then centrifuged for 1 min in an MSE Micro Centaur. Incubating the extract in normal goat serum and then protein A-Sepharose had no effect on the extract's activity; the extract used in the experiment shown was treated in this way. Anti-PDGF antibodies inhibited the activity of the extract to the same extent when they were added directly (85 μg ml^{-1}) to the cultures (data not shown). The results are expressed as means ± s.e.m. of triplicate cultures.

Our present and previous results provide compelling evidence that PDGF, secreted by type-1 astrocytes, regulates both the proliferation and the timing of differentiation of O-2A progenitor cells *in vitro*. To summarize the evidence: (1) PDGF stimulates the proliferation of O-2A progenitor cells and prevents them from differentiating prematurely into oligodendrocytes in culture (refs 9, 10, and this study); (2) cultures of type-1 astrocytes purified from cerebral cortex make both PDGF and messenger RNA encoding PDGF A chains[9]; (3) when ACM is fractionated by gel filtration, the mitogenic activity for O-2A progenitor cells is found in the same fractions as radiolabelled PDGF[9]; (4) anti-PDGF antibodies inhibit the ability of ACM and ONCM to stimulate O-2A progenitor cell proliferation *in vitro* (ref. 9, 10 and this study) and to restore the normal *in vivo* timing to oligodendrocyte differentiation (this study).

Table 4 Clonal analysis of oligodendrocyte differentiation in optic nerve cell cultures from 7-day-old rats stimulated by PDGF

	Number of cell divisions:				
Clones	0	1	2	3	4
a,b,c	1P	2ol			
d,e,f	1P	2P	4ol		
g	1P	2P	4P	8ol	
h	1P	2P	3P,1M	6ol	
i	1P	2P	4P	8P	16ol
j	1P	2P	4P	5P,1ol,1M	11ol

Cells from post-natal day 7 (P7) S/D rats were prepared and studied by time-lapse microcinematography as described[16], except that the cells were maintained in human PDGF (10 ng ml^{-1}) instead of ACM. Proliferation and differentiation of cells in microscopic fields containing 2 to 4 O-2A progenitor cells were followed for one week. The O-2A progenitor cells were identified by their characteristic bipolar morphology[15,16] and migratory behaviour[16], while oligodendrocytes were identified by their multipolar morphology[10,15,16] and lack of motility[16]. As shown in the table, oligodendrocyte differentiation occurred after one to four cell divisions in the ten clones studied, although in every experiment there were still dividing O-2A progenitor cells (belonging to other clones) present in the field at the end of filming. P, O-2A progenitor cells; ol, oligodendrocytes; M, cells that migrated out of the microscopic field and could no longer be followed.

Previous studies have suggested that ACM regulates the timing of oligodendrocyte development *in vitro* by keeping O-2A progenitor cells dividing until an intrinsic clock in the progenitor cell initiates the process that leads to oligodendrocyte differentiation[7,15]. The results of these studies were consistent with the possibility that the clock operates by setting a maximum number of divisions a progenitor cell and its progeny can undergo before differentiating. To test this possibility further we cultured optic nerve cells from seven-day-old rats in the presence of PDGF and followed the fate of individual O-2A progenitor cells and their progeny by time-lapse microcinematography. As described previously, O-2A progenitor cells and oligodendrocytes could be easily recognized as motile bipolar cells and immotile multipolar cells, respectively[16]. In nine of the ten clones studied, all of the descendants of a single progenitor cell differentiated together into non-dividing oligodendrocytes after the same number of cell divisions (Table 4); in the other clone (clone j in Table 4), the progenitor cells differentiated within one cell division of one another. These findings are consistent with previous single-cell experiments and the 'mitotic clock' hypothesis[15] but do not exclude other timing mechanisms.

Whatever the timing mechanism, it is clear that oligodendrocyte differentiation is associated with withdrawal from the cell cycle[4,5,8]. The relationship between the two processes, however, is uncertain. The clock in the progenitor cell might primarily control the onset of oligodendrocyte differentiation, with the cessation of proliferation following as a consequence. Alternatively, the clock might primarily control the onset of unresponsiveness to PDGF, with oligodendrocyte differentiation following as a consequence of withdrawal from the cell cycle. We favour the second possibility as it would most simply explain why O-2A progenitor cells differentiate prematurely when deprived of PDGF (see Table 2).

Whereas oligodendrocyte differentiation seems to be the constitutive pathway of O-2A progenitor cell development, which is automatically triggered when a progenitor cell is deprived of signals from other cells[17] or when the instrinsic timer reaches the appropriate point[7,15], type-2 astrocyte differentiation seems to be induced by a protein signal that greatly increases in concentration in the optic nerve after the first postnatal week[18]. While the mechanisms that control the timing and direction of O-2A progenitor cell differentiation *in vitro* are beginning to emerge, it remains to show that the same mechanisms operate *in vivo*.

We thank E. Raines, R. Ross, P. Stroobant and M. Waterfield for gifts of PDGF, C.-H. Heldin for a gift of anti-PDGF antibodies, P. Riddle for help with the time-lapse microcinematography and A. Mudge for helpful comments on the manuscript. L.E.L. was supported by a Fellowship from the National Multiple Sclerosis Society of the United States and W.D.R. by an MRC Project Grant.

Received 4 April; accepted 29 April 1988.

1. Raff, M. C., Miller, R. H. & Noble, M. *Nature* **303**, 390-396 (1983).
2. Raff, M. C., Abney, E. R., Cohen, J., Lindsay, R. & Noble, M. *J. Neurosci.* **3**, 1289-1300 (1983).
3. Miller, R. H., David, S., Patel, R., Abney, E. R. & Raff, M. C. *Devl Biol.* **111**, 35-41 (1985).
4. Skoff, R. P., Price, D. L. & Stocks, A. *J. comp. Neurol.* **169**, 291-312 (1976).
5. Skoff, R., Price, D. & Stocks, A. *J. comp. Neurol.* **169**, 313-333 (1976).
6. Raff, M. C., Williams, B. P. & Miller, R. H. *EMBO J.* **8**, 1857-1864 (1984).
7. Raff, M. C., Abney, E. R. & Fok-Seang, J. *Cell* **42**, 61-69 (1985).
8. Noble, M. & Murray, K. *EMBO J.* **3**, 2243-2247 (1984).
9. Richardson, W. D., Pringle, N., Mosley, M., Westermark, B. & Dubois-Dalcq, M. *Cell* **53**, 309-319 (1988).
10. Noble, M., Murray, K., Stroobant, P., Waterfield, M. D & Riddle, P. *Nature* **333**, 560-562 (1988).
11. Raff, M. C. *et al. Nature* **274**, 813-816 (1978).
12. Raines, B. & Ross, R. *J. biol. chem.* **257**, 5154-5160 (1982).
13. Stroobant, P. & Waterfield, M. *EMBO J.* **3**, 2963-2967 (1984).
14. Heldin, C.-H., Westermark, B. & Wasteson, A. *Expl Cell Res.* **136**, 255-261 (1981).
15. Temple, S. & Raff, M. C. *Cell* **44**, 773-779 (1986).
16. Small, R. K., Riddle, P. & Noble, M. *Nature* **329**, 155-157 (1987).
17. Temple, S. & Raff, M. C. *Nature* **313**, 223-225 (1985).
18. Hughes, S. & Raff, M. C. *Development* **101**, 157-167 (1987).
19. Ranscht, B., Clapshaw, P. A., Price, J., Noble, M. & Seifert, W. *Proc. natn. Acad. Sci. U.S.A.* **79**, 2709-2713 (1982).
20. Eisenbarth, G. S., Walsh, F. S. & Nirenberg, M. *Proc. natn. Acad. Sci. U.S.A.* **76**, 4913-4917 (1979).
21. Janzer, R. C. & Raff, M. C. *Nature* **325**, 253-257 (1987).
22. Yong, V. W. & Kim, S. U. *J. Neurosci. Meth.* **21**, 9-16 (1987).
23. Magaud, J.-P. and Mason, D. Y. *J. immunol. Meth.* (in the press).

The major Fc receptor in blood has a phosphatidylinositol anchor and is deficient in paroxysmal nocturnal haemoglobinuria

Periasamy Selvaraj, Wendall F. Rosse*, Robert Silber† & Timothy A. Springer

Laboratory of Membrane Immunochemistry,
Dana Farber Cancer Institute, Pathology Department,
Harvard Medical School, Boston, Massachusetts 02115, USA
* Department of Medicine, Duke University Medical Center,
Durham, North Carolina 27710, USA
† Department of Medicine, New York University Medical Center,
New York 10016, USA

Fc receptors on phagocytic cells in the blood mediate binding and clearance of immune complexes, phagocytosis of antibody-opsonized microorganisms, and potently trigger effector functions, including superoxide anion production and antibody-dependent cellular cytotoxicity. The Fc receptor type III (FcγR III, CD16), present in 135,000 sites per cell 1 on neutrophils and accounting for most of FcR in blood, unexpectedly has a phosphatidylinositol glycan (PIG) membrane anchor. Deficiency of FcγR III is observed in paroxysmal nocturnal haemoglobinuria (PNH), an acquired abnormality of haematopoietic cells[2] affecting PIG tail biosynthesis or attachment[3], and is probably responsible for circulating immune complexes[4] and susceptibility to bacterial infections associated with this disease[5]. Although a growing number of eukaryotic cell-surface proteins with PIG-tails are being described[6,7], none has thus far been implicated in receptor-mediated endocytosis or in triggering of cell-mediated killing. Our findings on the FcγR III raise the question of how a PIG-tailed protein important in immune complex clearance *in vivo*[8,9] and in antibody-dependent killing[10] mediates ligand internalization and cytotoxicity. Together with our results, previous functional studies on FcγR III and FcγR II[11,12] suggest that these two receptors may cooperate and that the type of membrane anchor is an important mechanism whereby the functional capacity of surface receptors can be regulated.

Three different types of FcγR have been distinguished in humans using monoclonal antibodies[13] (mAb). FcγR III (CD16) of relative molecular mass (M_r) 50-70,000 (50-70K) is found on neutrophils, large granular lymphocytes, and macrophages, but not on monocytes. FcγR III was first identified with a mAb (3G8) that blocked immune complex binding to neutrophils[1] and subsequently with other mAb of the CD16 cluster[14]. FcγR II (CDw32) is a 40K receptor on neutrophils, monocytes, eosinophils, platelets and B cells[13]. FcγR I is a 72K protein and is found on monocytes[13]. FcγR III and FcγR II have low affinity for monomeric IgG and thus preferentially bind immune complexes by multiple receptor-ligand interactions, whereas FcγR I is sufficiently high affinity to bind monomeric IgG.

Our first evidence that FcγR III is anchored by PIG came from studies on leukocytes from patients with paroxysmal nocturnal haemoglobinuria (PNH). PNH is an acquired defect of haematopoietic precursor cells in either the biosynthesis or the attachment of the PIG tail and may affect clonal progeny in the erythroid, monocytic, granulocytic, and thrombocytic lineages[2,15,16]. Previous studies on erythrocytes and leukocytes from PNH patients have demonstrated a selective deficiency of PIG-tailed proteins, including decay accelerating factor (DAF), acetylcholinesterase, alkaline phosphatase and the PIG-anchored form of lymphocyte function-associated antigen 3 (LFA-3), (refs 3, 6 and 7). The deficiency of DAF accounts for susceptibility of erythrocytes to complement-mediated lysis in PNH. However none of these previously identified deficiencies can explain the occurrence of circulating immune complexes[4] and the 20% and 50% of mortalities caused by bacterial infections and thrombosis respectively[5].

Quantitation of FcγR III expression using immunofluorescence flow cytometry show that it is markedly deficient on PNH neutrophils (Fig. 1*a*). This deficiency was found in all six patients studied (D.E., S.B., J.M., J.E., B.I., C.G.) and results with five different CD16 (FcγR III) mAb were identical. Some patients such as J.E. (Fig. 1*a*, curve 3) showed normal as well as deficient granulocyte clones. Patients showed consistent variation in the extent of deficiency in the abnormal clone. The amount of FcγR III expression on affected cells ranged from 2% (patient D.E.) to 19% (patient J.E.) averaging 7% of normal, perhaps reflecting the degree of penetrance of the acquired defect in PNH. In all cases, deficiency of FcγRIII paralleled deficiency of DAF. In contrast, deficient neutrophils expressed normal levels of HLA-A,B, LFA-1, Mac-1 and FcγR II (CDw32) (Fig. 1*a*). PNH monocytes showed normal expression of FcγR I and II, although they were deficient in DAF (Fig. 1*b*). These results suggested that the neutrophil FcγR III has a PIG tail, whereas the FcγR I and FcγR II do not.

PIG-anchored proteins can be specifically cleaved from cell surfaces with phosphatidylinositol-specific phospholipase C[3,6,7] (PIPLC). We therefore investigated the susceptibility of Fc receptors to PIPLC (Table 1). PIPLC released 75-84% of the cell surface FcγR III and DAF from healthy granulocytes, while FcγR II, HLA-A,B and LFA-1 were unaffected. On monocytes, PIPLC released 84% of DAF whereas it had no effect on FcγR II, FcγR I, HLA-A,B and LFA-1. Results with PIPLC prepared from *S. aureus* and *B. thuringiensis* were identical and show that FcγR III on neutrophils, but not FcγR II on the same cells or FcγR I and II on monocytes, have PIG anchors. Lack of a PIG anchor on FcγR II is consistent with the prediction (from cDNA sequence[17]) that it possesses a transmembrane domain and a 76 residue hydrophilic cytoplasmic tail.

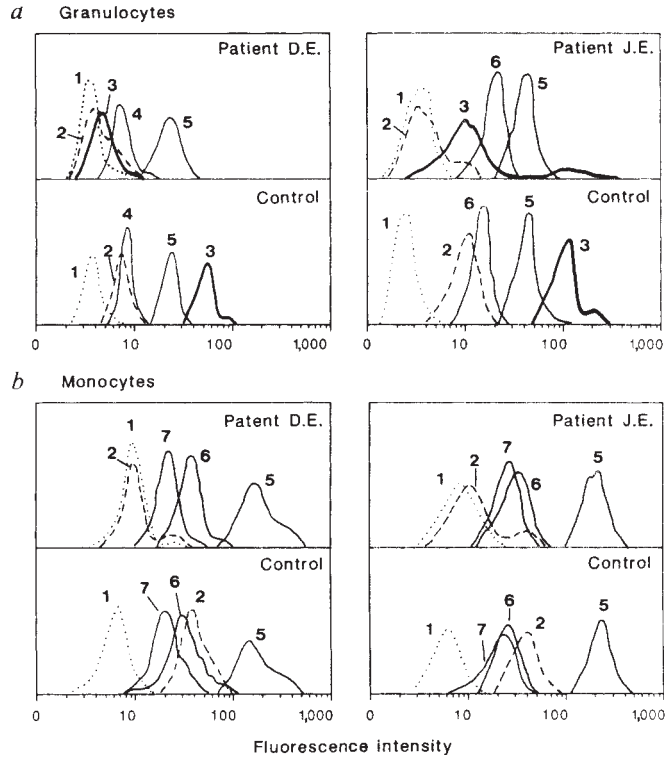


Fig. 1 Immunofluorescence flow cytometry analysis of granulocytes (*a*) and monocytes (*b*) from PNH patients and normal controls. PNH patient D.E. had type III erythrocytes and patient J.E. had type III and type I erythrocytes. 1, ... X63 control; 2, - - - - DAF; 3, —— FcγR III (3G8); 4, LFA-1; 5, HLA-A, B; 6, FcγR II; 7, FcγR I (4-7, ——).

Methods. Blood samples were collected from PNH patients and from healthy individuals in acid citrate dextrose (ACD) as anticoagulant. Leukocyte-rich plasma was separated from erythrocytes by 6% dextran T-500 (Pharmacia) sedimentation, overlaid on Histopaque-1077 (Sigma Chemical Co.) and centrifuged at 1,000 *g* for 25 min. Mononuclear cells were recovered from the interface and granulocytes were recovered from the pellet with some erythrocytes. The contaminating erythrocytes were removed by hypotonic lysis with water for 20 s. For some experiments granulocytes were pretreated with fMet-Leu-Phe as described[26]. Cells were washed five times, stained with hybridoma culture supernatants diluted two fold or ascites diluted 400-fold with fetal calf serum (FCS)/RPMI 1640 followed by FITC conjugated goat anti-mouse IgG or IgM (Zymed), and subjected to flow cytometry as previously described[26]. Monoclonal antibodies have been previously referenced: 1A10 to DAF (ref. 16), 3G8 to FcγR III (ref. 1); W6/32 to HLA-A,B, TS1/22 to LFA-1, LM2/1 to Mac-1 (ref. 26), and 32 to FcγR I (ref. 27) or were from the third international workshop on human leukocyte differentiation antigens[14]: 2E1 and CIKM5 to FcγR II, and CLBFc gran-1, MG38, G2G4, BW243/41 and G7E11 to FcγR III. P3X63 myeloma IgG and OCH.217 CD2 IgM antibody were used as non-binding controls for IgG and IgM staining respectively.

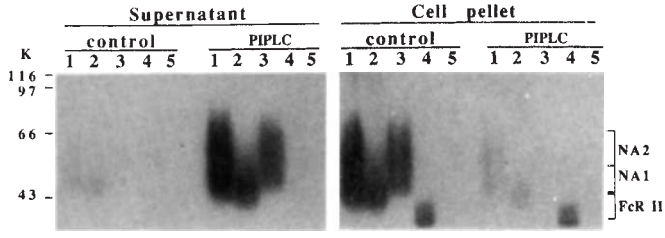


Fig. 2 SDS-PAGE of FcγR III after PIPLC treatment. FcγR was precipitated from the supernatant or cell lysate after treatment of ^{125}I-labelled neutrophils, with or without PIPLC. Lane 1, FcγR III precipitated by CLBFc gran-1; lane 2, NA1 allotype of FcγR III precipitated by MG38; lane 3, NA2 allotype of FcγR III precipitated by CLBFc gran-1 after preclearing with MG38; lane 4, FcγR II precipitated by CIKM5; lane 5, X63 control precipitate, Size markers are indicated to the left.

Method. Granulocytes were surface iodinated with ^{125}I using 1,3,4,6-tetrachloro-3α,6α-diphenyl glycoluril[28]. Aliquots of the iodinated cells were washed three times with HBSS/HEPES, *p*H 7.4, and treated with *Bacillus thuringiensis* PIPLC as described under Table 1. Cells were centrifuged at 340 *g*. The resultant supernatant was further centrifuged at 100,000 *g* and saved. The cell pellet was washed with 10% FCS/RPMI and lysed in 1% Triton X-100 in 50 mM Tris/HCl buffer, *p*H 8.0, containing 0.02% azide, 1% bovine haemoglobin, 5 mM iodoacetamide, 1 mM PMSF, 1 mM diisopropyl fluorophosphate and 1% aprotinin. The lysate and 100,000 *g* supernatant were dialysed overnight against 10 mM Tris, 145 mM NaCl and 0.02% azide. Samples (20 μl) were incubated at 4 °C and 18 μl ascites (diluted 1:50), 5 μl rabbit anti-mouse IgG (Zymed) and 30 μl protein A-agarose (Genzyme) were added sequentially at 4 h intervals. The resulting immunoprecipitates were washed three times with lysis buffer (diluted 1:10), once with 50 mM Tris/HCl, *p*H 8.0 and analysed by SDS PAGE (10% gels) under non-reducing conditions The gel was dried and autoradiographed; only the relevant portion is shown.

FcγR III is polymorphic[18]. The alloantigens NA1 and NA2 are distinguished by human alloantisera, by size in SDS-PAGE, and by certain CD16 mAb. The results in Table 1 were obtained with NA1, NA1 homozygotes and NA1, NA2 heterozygotes. The amount of release with PIPLC was similar whether measured with MG38 mAb specific for NA1 or with four other CD16 mAb (CLB FcR gran-1, BW243/4, G2G4, and G7E11) which react with both alloantigens (Table 1). These CD16 MAb have been placed in three distinct groups by Tetteroo *et al.*[14], based on intensity of staining or differential reactivity with neutrophils and large granular lymphocytes.

FcγR III bearing NA1 has higher mobility than that of NA2 (refs 14, 18). Sequential immunoprecipitation using a NA1-specific allotypic mAb and a monotypic mAb showed that both NA1 and NA2 allotypes were released by PIPLC treatment (Fig. 2 lanes 2 and 3). FcγR II (Fig. 2 lane 4) and LFA-1 and Mac-1 (data not shown) were not released. A small amount of FcγR III was released in the absence of PIPLC, possibly due to the action of endogenous phospholipases. The amount of spontaneous release appeared to be higher with neutrophils which had been prestimulated with fMet-Leu-Phe (not shown). Soluble FcγR III has been detected in plasma[19].

We conclude that the neutrophil FcγR III has a PIG anchor. FcγR III on neutrophils accounts for ~50% of all Fcγ receptors in the blood and 80% of those found on phagocytic cells (Table 2). Its deficiency is likely to underlie the presence of circulating immune complexes in PNH[4] and to be related to the susceptibility of PNH patients to bacterial infections[5]. Administration of mAb or Fab to the FcγR III profoundly inhibits immune complex clearance *in vivo*[8,9], further supporting its important role. In addition to that on neutrophils, FcγR III in spleen red pulp and liver Kuppfer cells may be important in *in vivo* clearance[9]. The nature of the anchor on these cells remains to be determined.

The neural cell adhesion molecule (NCAM) and LFA-3 exist in both PIG- and polypeptide-anchored forms[7]. Although we cannot rule out a polypeptide-anchored form of FcγR III on neutrophils, 98% deficiency in PNH patients D.E. (Fig. 1) and J.M. (data not shown) suggest that if present at all, it would constitute ≤2% of the total receptor. Incomplete release of PIG-tailed proteins with PIPLC is always observed, most probably due to acylation at the 2-OH of inositol[7], and does not indicate the existence of a polypeptide-anchored form.

Table 1 Effect of PIPLC treatment on binding of FcγR mAbs to granulocytes and monocytes

	Fluorescence Intensity			
	Control	+PIPLC *S. aureus*	+PIPLC *B. thuringiensis*	% decrease (Mean ± s.d.)
Granulocytes				
DAF	2.1	0.7	0.3 ± 0.2	75 ± 9
FcγR III				
3G8	13.5	3.3	2.4 ± 0.7	78 ± 2
CLB FcR gran1	14.3	4.0	2.7 ± 0.9	77 ± 4
MG38	12.7	3.7	2.8 ± 0.9	75 ± 4
G2G4	22.6	—	3.2 ± 2.7	84 ± 14
G7E11	27.5	—	4.7 ± 3.0	83 ± 12
BW243/41	0.9	—	0.2 ± 0.12	76 ± 13
FcγR II	4.2	4.3	4.9 ± 0.26	0
LFA-1	2.1	2.3	2.3 ± 0.1	0
Mac-1	8.9	9.4	8.9 ± 0.1	0
HLA-A, B	10.7	11.6	11.4 ± 0.7	0
Monocytes				
DAF	10.6	—	1.7	84
FcγR I	5.4	—	6.0	0
FcγR II	7.8	—	8.5	0
HLA-A, B	62.8	—	68.1	0

Cells (5-10 × 10^6 per ml) were resuspended in HBSS/HEPES *p*H 7.4 containing 1 mg ml^{-1} ovalbumin; treated with *S. aureus* (10 μg ml^{-1}) or *B. thuringiensis* (0.6 mmol min^{-1} ml^{-1}) PIPLC for 45 min at 37 °C and analysed by flow cytometry as described in Fig. 1. The antibody staining on control cells which were treated in the same way but without PIPLC was taken as 100%. For comparative purposes the control values were normalized. Monocytes and granulocytes were from different donors and were analysed at different times. Among FcγR III mAbs MG38 reacts only with NA1, whereas others react with both NA1 and NA2. The averaged results with *B. thuringiensis* PIPLC were from two homozygous (NA1, NA1) and a heterozygous (NA1, NA2) donor. The specific activity of *B. thuringiensis* PIPLC was 100-fold higher than *S. aureus* PIPLC (M. Low, personal communication).

An important question is whether the PIG anchor, with its diacylglycerol moiety present in only the outer leaflet of the membrane bilayer, can deliver signals. Although mAb-induced cross-linking of Thy-1 and Ly-6 on T cells can trigger functions such as mitogenesis[20,21], the natural ligands of these PIG-tailed proteins are unknown, and it is not clear whether Thy-1 or Ly-6 trigger physiologically. Because Fc receptors are potent physiological triggers of neutrophil and monocyte function, it is of considerable interest that the most abundant of these is PIG-tailed. Monoclonal antibodies and Fab fragments to FcRγ III have previously been found to block binding and hence internalization of soluble immune complexes and antibody-opsonized erythrocytes[1] to neutrophils, and to inhibit antibody-dependent killing of target cells[10]. This is not direct evidence for triggering by FcγR III, however, as both FcγR III and FcγR II will be engaged by immune complexes and thus may act synergistically. The roles of FcγR III and FcγR II in triggering have been partly resolved by antibody-redirected killing. Resting or IFN-γ-stimulated neutrophils kill chicken erythrocytes in the presence of anti-FcγR III Fab conjugated to anti-chicken erythrocyte Fab[11], indicating that a PIG-tailed receptor can trigger. But FcγR I induced by IFN-γ on neutrophils is a better trigger, despite its lower density. Assays on hybridoma cells bearing anti-FcR surface Ig show that FcγR III does not trigger neutrophil killing of tumour cells, whereas FcγR II present at a four-fold lower density does[12]. Studies on neutrophil superoxide anion generation suggest that FcγR III is required for binding of immune complexes and FcγR II has an important role in triggering[1,22].

Table 2 Fc receptors in blood

		Sites per cell* $\times 10^{-3}$	Blood concentration (nM)†
Neutrophil	FcγR II	31	0.22
	FcγR III	135	1.0
Monocyte	FcγR I	20	0.01
	FcγR II	36	0.014
B cell	FcγR II	38	0.02
Platelet	FcγR II	1.4	0.7

* Sites per cell from refs 1, 13, 23 and 24 and our own unpublished results.

† Calculated from cell concentrations in blood[25].

Thus functional studies together with our results suggest that a physiologically relevant PIG-anchored receptor, FcγR III, can trigger at least one type of neutrophil function (killing of erythrocytes[11]) but that it triggers less efficiently than the protein-anchored FcγR I and FcγR II (refs 11, 12). Although we cannot rule out influences of the extracellular domains of these receptors, the findings suggest that the type of membrane anchor may be an important mechanism for regulating the functional capacity of surface receptors. The FcγR III is present in more copies per cell (135,000) than any other known receptor on a circulating cell. Immune complexes must be removed from the circulation without causing disseminated neutrophil activation; otherwise severe immunopathological reactions ensue. The PIG anchor may allow efficient ligand binding and internalization to be coupled with relatively weak signalling.

The FcγR III is coexpressed on neutrophils with the FcγR II. The ability of mAb to either the FcγR II or FcγR III to inhibit neutrophil functions[1,10,22] suggests that these low affinity receptors may synergize in interactions with multivalent immune complexes. The high density of the FcγR III and its predicted fast mobility in the membrane bilayer[6,7] may suit it for capture of immune complexes, while the FcγR II may have the major role in triggering.

This work was supported by grants from the NIH. We thank Dr M. Low for providing PIPLC, Drs J. Unkeless and M. Fanger for mAb, Dr A. Nicholson-Weller for blood from patient B.I. and P. Lopez for flow cytometry.

Received 8 March; accepted 11 April 1988.

1. Fleit, H. B., Wright, S. D. & Unkeless, J. C. *Proc. natn. Acad. Sci. U.S.A.* **79**, 3275-3279 (1982).
2. Rosse, W. F. & Parker, C. J. *Clin. Haemat.* **14**, 105-125 (1985).
3. Low, M. G. *Biochem. J.* **244**, 1-13 (1987).
4. Villaescusa, R. *Acta haemat.* **68**, 136-141 (1982).
5. Crosby, W. H. *Blood* **8**, 769-812 (1953).
6. Low, M. G. & Saltiel, A. R. *Science* **239**, 268-275 (1988).
7. Ferguson, M. A. J. & Williams, A. F. *A. Rev. Biochem.* (in the press).
8. Clarkson, S. B., Bussel, J. B., Kimberley, R. P., Valinsky, J. E., Nachman, R. L. & Unkeless, J. C. *N. Engl. J. Med.* **314**, 1236-129 (1986).
9. Clarkson, S. B. *et al. J. exp. Med.* **164**, 474-489 (1986).
10. Perussia, B. et al. J. Immun. **130**, 2142-2148 (1983).
11. Shen, L., Gure, M. & Fanger, M. W. *J. Immun.* **139**, 534-538 (1987).
12. Graziano, R. F. & Fanger, M. W. *J. Immun.* **139**, 3536-3541 (1987).
13. Anderson, C. L. & Looney, R. J. *Immun. Today* **7**, 264-266 (1986).
14. Tetteroo, P. A. T., Van Der Schoot, C. E., Visser, F. J., Bos, M. J. E. & Von deme Borne, A. E. G. in *Leucocyte typing III* (eds McMichael, A. J. *et al.*) 702-706 (Oxford University Press, Oxford, 1987).
15. Nicholson-Weller, A., Spicer, D. B. & Austen, K. F. *New Engl. J. Med.* **312**, 1091-1097 (1985).
16. Kinoshita, T., Medof, M. E., Silber, R. & Nussenzweig, V. *J. exp. Med.* **162**, 75-92 (1985).
17. Stuart, S. G. *et al. J. exp. Med.* **166**, 1668-1684 (1987).
18. Werner, G. *et al* in *Leukocyte Typing II* (eds Reinherz, E. L., Haynes, B. F., Nadler, L. M. & Bernstein, I. D.) 109-121 (Springer, New York, 1986).
19. Khayat, D. *et al. J. immunol. Meth.* **100**, 235-241 (1987).
20. Reiser, H. *et al. Cell* **47**, 365-370 (1986).
21. Kroczek, R. A., Gunter, K. C., Seligman, B. & Shevach, E. M. *J. Immun.* **136**, 4379-4384 (1986).
22. Looney, R. J. *et al. J. exp. Med.* **163**, 826-836 (1986).
23. Karas, S. P., Rosse, W. F. & Kurlander, R. J. *Blood* **60**, 1277-1282 (1982).
24. Vaughan, M., Taylor, M. & Mohankumar, T. *J. Immun.* **135**, 4059-4065 (1985).
25. Williams, W. J., Beutler, E., Erselev, A. J. & Lichtman, M. N. *Hematology* (McGraw-Hill, New York, 1983).
26. Miller, L. J., Bainton, D. F., Borregaard, N. & Springer, T. A. *J. clin. Invest.* **80**, 535-544 (1987).
27. Anderson, C. L. *et al. J. immun.* **261**, 12856-12864 (1986).
28. Fraker, P. J. & Speck, J. C. *Biochem. biophys. Res. Commun.* **80**, 849-857 (1978).
29. Laemmli, U. K. *Nature* **227**, 680-685 (1970).

The Fcγ receptor of natural killer cells is a phospholipid-linked membrane protein

David Simmons & Brian Seed

Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

Three types of receptor for the Fc (constant) region of human immunoglobulin G have been described[1,2]; FcRI, a high-affinity ($K_a \approx 10^8$ M^{-1}) receptor expressed on monocytes[3–5]; FcRII (CD32), a low-affinity ($K_a \approx 10^6$ M^{-1}) receptor expressed on B cells, granulocytes, macrophages and platelets[6–9]; and FcRIII (CD16, FcR_{lo}), a low-affinity receptor expressed on macrophages, neutrophils, eosinophils, natural killer cells and a subset of T cells believed to comprise the suppresssor cells[10–13]. Anti-CD16 antibodies block natural killer-cell mediated antibody dependent cellular cytotoxicity (ADCC)[14,15]. Binding of aggregated IgG to CD16 on natural killer cells leads to the expression of lymphocyte activation antigens, mediator release, morphological changes and lytic activity[16–18]. We report here the isolation of a complementary DNA clone encoding CD16 determinants which gave rise to IgG binding of the expected affinity and subtype specificity in COS cells, and which proved to encode a phospholipid anchored protein. A single messenger RNA transcript was found in all positive RNA samples, and N-glycanase treatment showed the form found in COS cells was identical to the form present on peripheral blood mononuclear cells (PBMCs). We also show that CD16 is most closely related to the α-form of the murine IgG 2b/1 receptor and propose that extracellular contacts mediate the signal initiated by IgG binding.

A cDNA library was constructed in the expression vector CDM8 (ref. 19) from the poly(A)$^+$ fraction of RNA prepared from human placenta (a rich source of monocyte transcripts[9]). The library was introduced into COS cells by DEAE dextran transfection and cells expressing CD16 were recovered 48 hours later by panning with a panel of anti-CD16 monoclonal antibodies (mAbs) from the Third Leukocyte Typing Workshop[20] (BW 209/2, GRM1, CLB/Fcgran I and MG38). Episomal DNA was recovered from the adherent cells, amplified in *Escherichia coli* and re-introduced into COS cells by spheroplast fusion. After three such rounds of expression, panning and rescue, DNA was prepared from individual *E. coli* colonies and used to transfect COS cells, allowing CD16 expression to be determined by indirect immunofluorescence 48 hours later. Four out of eight individual transfectants scored positive, each having an insert of 1 kilobase pair (kb). A single clone, designated pCD16, was selected for further study.

COS cells transfected with pCD16 gave positive surface immunofluorescence reactions at nM concentrations with all four mAbs and a further eight anti-CD16 mAbs[20] (FC-1, FC-2, VI-FcR2, BW243/41, Leu11 (L23), Leu11c/B73.1, GO22b and MG12) (data not shown).

Scatchard analysis of the binding of radiolabelled IgG1 to transfected COS cells showed the presence of $2\text{–}4 \times 10^6$ sites per cell with an association constant (K_a) $\approx 5 \times 10^5$ M^{-1} (Fig. 1*a*). The isotype specificity of the transfected cells was determined by ligand displacement titrations with both murine and human IgGs[9,21,22]. In agreement with *in vivo* data[23,24], the pCD16-encoded receptor preferentially bound murine IgG2a and 3 ($K_a \approx 3 \times 10^5$ M^{-1}), and had no detectable affinity for IgG2b ($\ll 10^5$ M^{-1}; Fig. 1*b*). But the transfected cells did show intermediate affinity for murine IgG1 ($K_a \approx 2 \times 10^5$ M^{-1}). The rank order for binding of human IgGs (IgG1.3 > 2, 4; Fig. 1*c*) similarly agreed with *in vivo* data[25,26]. The transfected cells had no detectable affinity for other classes of human immunoglobin (A, D, E, or M) as assayed by indirect immunofluorescence (data not shown).

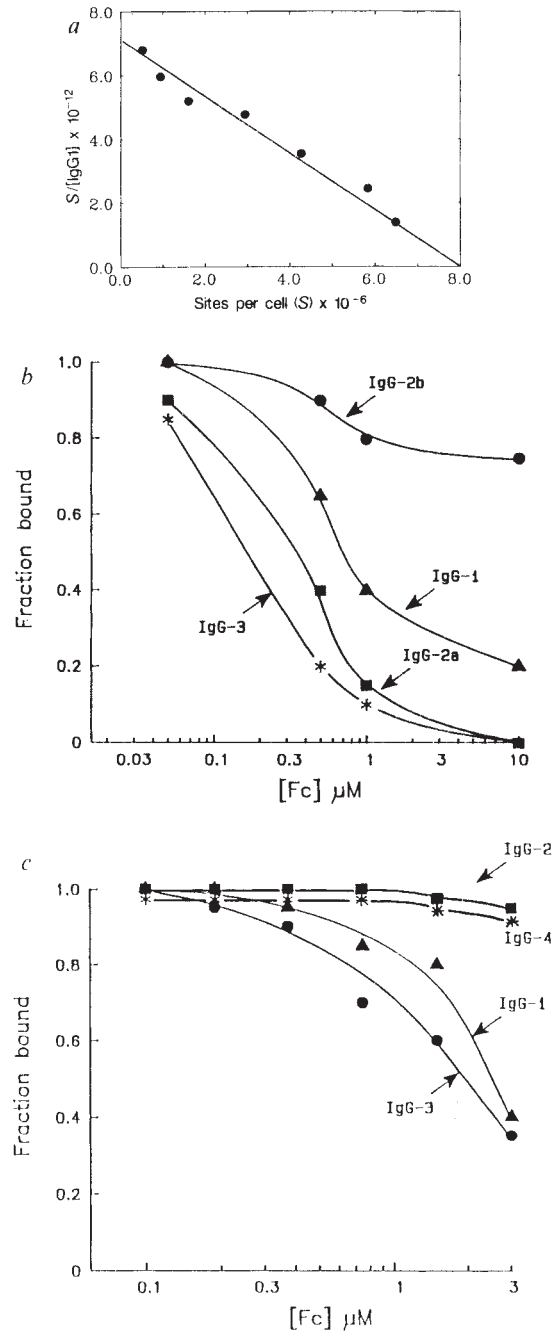


Fig. 1 Ligand binding analysis of CD16. *a*, Scatchard analysis of the binding of ^{125}I-IgG1 to transfected COS cells was performed as described in ref. 9. Ligand was incubated with transfected cells for 30 min at 4 °C in phosphate-buffered saline (PBS), 5% non-fat milk, 0.2% NaN_3. Bound and free fractions were separated by centrifugation through a 3 : 2 mixture of dibutyl : dioctyl phthalates. Data were corrected for the contribution of nonspecific binding by subtraction of label (c.p.m.) bound in the presence of 2.5×10^{-4} M total IgG. *b* and *c*, Ligand-displacement analysis of binding of IgG subtypes. The fraction of ^{125}I-labelled IgG1 tracer bound in the presence of non-radioactive IgGs is plotted as a function of the Fc concentration of the competitor. *b*, Murine IgG subtypes 1, 2a, 2b and 3. *c*, Human subtypes 1, 2, 3 and 4. Binding of ligand and separation of bound and free fractions was carried out as in *a*.

RNA blot hybridization showed a single species of ≈2 kb present in placenta (Fig. 2*a*), LAK cells (lymphokine activated killer cells) and short-term, bulk cultures of PBMCs enriched for CD16$^+$/HNK-1$^+$ (natural killer) cells (Fig. 2*c*). CD16 transcripts were absent from all other cell lines tested including

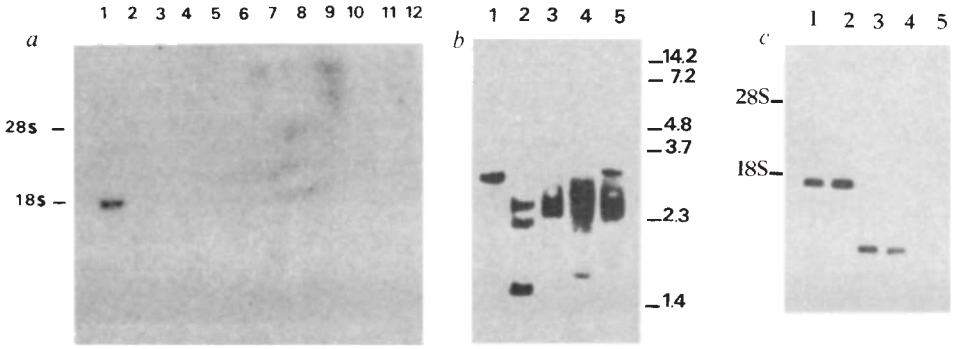
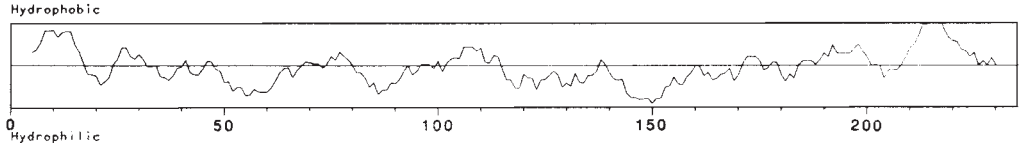


Fig. 2 Expression and structure of 4he CD16 gene. *a*, RNA blot analysis of CD16 transcripts. 20 μg of total RNA from different sources was fractionated on a 1% agarose/formaldehyde gel, transferred to nylon and hybridized with a ^{32}P-labelled single-stranded form of pCD16. Lane 1, placenta; lane 2, K562 (erythroleukaemia); lane 3, HEL (erythroleukaemia); lane 4, U937 (promonocytic leukaemia); lane 5, HL60 (promyelocytic leukaemia); lane 6, HUT102 (T-cell leukaemia); lane 7, Molt 4 (T-cell leukaemia); lane 8, Jurkat (T-cell leukaemia); lane 9, Raji (Burkitt lymphoma); lane 10, Daudi (Burkitt lymphoma); lane 11, JY (lymphoblastoid cell line); lane 12, HepG2 (hepatoblastoma). Positions of 18S and 28S rRNA are indicated on the left. *b*, Genomic DNA blot hybridization analysis. 20 μg of human placenta genomic DNA were digested with restriction enzymes, electrophoresed, transferred to nitrocellulose and hybridized with the CD16 cDNA. The enzymes used were; lane 1, *Eco*RI; lane 2, *Hin*d III; lane 3, *Bam*HI; lane 4, *Pst*I; lane 5, *Dra*I. Molecular sizes in kilobase pairs are on the right. *c*, S1 analysis of CD16 transcripts. Poly(A)$^+$ RNA (5 μg) from LAK cells and PBMC cultures enriched for CD16$^+$ NK cells were hybridized with a single-stranded form of the CD16 cDNA at 50 °C for 18 h in formamide buffer[34], digested with S1 nuclease (1 unit per μg RNA) and precipitated. The protected fragments were run on a 1% agarose/formaldehyde gel, transferred to nylon and hybridized with the CD16 cDNA insert as a probe. Lanes 1 and 2, untreated poly(A)$^+$ RNA from LAK cells (lane 1) and from PBMC cultures enriched for NK cells (lane 2). Lanes 3-5, samples hybridized with CD16 and digested with S1; LAK RNA (lane 3); NK enriched RNA (lane 4); control, *E. coli* transfer RNA (lane 5). Positions of 18S and 28S RNA are indicated on the right.

a

```
1   TCTTTGGTGACTTGTCCACTCCAGTGTGGCATCATGTGGCAGCTGCTCCTCCCAACTGCTCTGCTACTTCTAGTTTCAGCTGGCATGCGGACTGAAGATCTCCCAAAGGCTGTGGTGTTC 120
1                                   MetTrpGlnLeuLeuLeuProThrAlaLeuLeuLeuLeuValSerAlaGlyMetArgThrGluAspLeuProLysAlaValValPhe 29

121 CTGGAGCCTCAATGGTACAGCGTGCTTGAGAAGGACAGTGTGACTCTGAAGTGCCAGGGAGCCTACTCCCCTGAGGACAATTCCACACAGTGGTTTCACAATGAGAGCCTCATCTCAAGC 241
30  LeuGluProGlnTrpTyrSerValLeuGluLysAspSerValThrLeuLysCysGlnGlyAlaTyrSerProGluAspAsnSerThrGlnTrpPheHisAsnGluSerLeuIleSerSer 70
                                                                                   ---CHO---            ---CHO---

242 CAGGCCTCGAGCTACTTCATTGACGCTGCCACAGTCAACGACAGTGGAGAGTACAGGTGCCAGACAAACCTCTCCACCCTCAGTGACCCGGTGCAGCTAGAAGTCCATATCGGCTGGCTG 362
71  GlnAlaSerSerTyrPheIleAspAlaAlaThrValAsnAspSerGlyGluTyrArgCysGlnThrAsnLeuSerThrLeuSerAspProValGlnLeuGluValHisIleGlyTrpLeu 111
                                        ---CHO---                  ---CHO---

363 TTGCTCCAGGCCCCTCGGTGGGTGTTCAAGGAGGAAGACCCTATTCACCTGAGGTGTCACAGCTGGAAGAACACTGCTCTGCATAAGGTCACATATTTACAGAATGGCAAAGACAGGAAG 483
112 LeuLeuGlnAlaProArgTrpValPheLysGluGluAspProIleHisLeuArgCysHisSerTrpLysAsnThrAlaLeuHisLysValThrTyrLeuGlnAsnGlyLysAspArgLys 152

484 TATTTTCATCATAATTCTGACTTCCACATTCCAAAAGCCACACTCAAAGATAGCGGCTCCTACTTCTGCAGGGGGCTTGTTGGGAGTAAAAATGTGTCTTCAGAGACTGTGAACATCACC 604
153 TyrPheHisHisAsnSerAspPheHisIleProLysAlaThrLeuLysAspSerGlySerTyrPheCysArgGlyLeuValGlySerLysAsnValSerSerGluThrValAsnIleThr 193
                                                                                   ---CHO---            ---CHO---

605 ATCACTCAAGGTTTGGCAGTGTCAACCATCTCATCATTCTCTCCACCTGGGTACCAAGTCTCTTTCTGCTTGGTGATGGTACTCCTTTTTGCAGTGGACACAGGACTATATTTCTCTGTG 725
194 IleThrGlnGlyLeuAlaValSerThrIleSerSerPheSerProProGlyTyrGlnValSerPheCysLeuValMetValLeuLeuPheAlaValAspThrGlyLeuTyrPheSerVal 234
                                                                  ==========================================================

726 AAGACAAACATTTGAAGCTCAACAAGAGACTGGAAGGACCATAAACTTAAATGGAGAAAGGACCCTCAAGACAAATGACCCCCATCCCATGGGAGTAATAAGAGCAGTGGCAGCAGCATC 846
235 LysThrAsnIleEnd                                                                                                          238

847 TCTGAACATTTCTCTGGATTTGCAACCCCATCATCCTCAGGCCTCTC
```

b



c

```
Common          L P K A V V   L   P   W     V L     D   V T L [C] G       P       S T Q W F H N       I   S Q
MFCGRA      S A A L P K A V V K L D P P W I Q V L K E D M V T L M [C] E G T H N P G N S S T Q W F H N G R S I R S Q V Q A S
CD16        T E D L P K A V V F L E P Q W Y S V L E K D S V T L K [C] Q G A Y S P E D N S T Q W F H N E S L I S S Q A S S Y
CD32        A A A P P K A V L K L E P P W I N V L Q E D S V T L T [C] Q G A R S P E S D S I Q W F H N G N L I P T H T Q P S
Common          P K A V     L E P   W     V L     D S V T L   [C] Q G A   S P E     S   Q W F H N     L I

Common          A T V N D S G E Y R [C] Q           L S D P V   L   V       W L L L Q   P     V F   E       I   L R [C] H
MFCGRA      Y T F K A T V N D S G E Y R [C] Q W E Q † R L S D P V D L G V I S D W L L L Q T P Q R V F L E G E T I † L R [C] H
CD16        F I D A A T V N D S G E Y R [C] Q T N L S T L S D P V Q L E V H I G W L L L Q A P R W V F K E E D P I H L R [C] H
CD32        Y R F K A N N N D S G E Y T [C] Q T G Q T S L S D P V H L T V L S E W L V L Q T P H L E F Q E G E T I M L R [C] H
Common          A     N D S G E Y     [C] Q T         L S D P V   L   V     , W L   L Q   P       F   E       I   L R [C] H

Common      S W   N     L             N   K       Y   H     S   F   I P K A         S G   Y [C] G     G S         S
MFCGRA      S W R N K L L N R I S F F H N E K S V R Y H H Y K S N F S I P K A N H S H S G D Y Y [C] K G S L G S T Q H Q S K
CD16        S W K N T A L H K V T Y L Q N G K D R K Y F H H N S D F H I P K A T L K D S G S Y F [C] R G L V G S K N V S S E
CD32        S W K D K P L V K V T F F Q N G K S Q K F S R L D P T F S I P Q A N H S H S G D Y H [C] T G N I G Y T L F S S K
Common      S W K       L   K V T     Q N G K     K             F   I P   A         S G   Y [C] G     G         S S

Common        V   I T             S   I S           Y       F   L V M   L       L F A V D T G L Y F   V       N
MFCGRA      P V T I T V Q D P A T T S S I S L V W - - - Y H T A F S L V M C L - - - - F A V D T G L Y F Y V R R N L
CD16        T V N I T I T Q G L A V S T I S S F S P P G Y Q V S F C L V M V L - - - - F A V D T G L Y F S V K T N I *
CD32        P V T I T V - - - - Q V P S M G S S S P M G I I V A V V I A T A V A A I V A A V V A L I Y C R K K R I S
Common        V   I T             V         S   S P   G     V                       A V         Y       K
```

Fig. 3 Sequence of CD16 cDNA and similarity with murine and human Fc receptors. *a*, Nucleotide sequence of pCD16 cDNA insert. The nucleotide sequence (above) and translated open reading frame (below) are numbered. Sites of potential N-linked glycosylation are depicted -CHO- and the hydrophobic carboxyl terminus is underlined. Sequencing was by dideoxy-chain termination[35], using specific oligonucleotide primers. *b*, Hydropathicity profile[36] of the predicted amino-acid sequence. *c*, Alignment of CD16 with the α-form of the murine Fcγ2b/1 receptor[31] (MFCGRA, above), and CD32 (PC23[9], below), produced by the ALIGN program of the Protein Identification Resource (NBRF[30]). Conserved cysteine residues (two per immunoglobulin domain) are boxed.

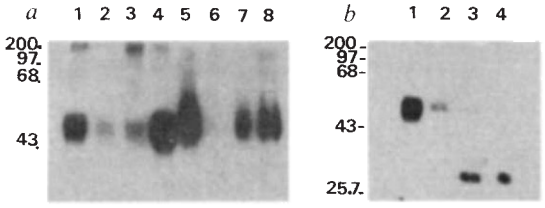


Fig. 4 Phospholipase treatment and immunoprecipitation of CD16 from ^{125}I-labelled cells. *a*, Lanes 1 to 4, pCD16 transfected COS cells. Lane 1, FC-1 immunoprecipitate from the cell fraction, or lane 2, the supernatant fraction, of untreated cells. Lane 3, FC-1 immunoprecipitate from the cell fraction, or lane 4, the supernatant fraction, of GPI-PLC treated samples. Lanes 5 to 8, nylon-wool-nonadherent PBMCs. Lane 5, FC-1 immunoprecipitate from the cell fraction, or lane 6, the supernatant fraction of untreated cells. Lane 7, FC-1 immunoprecipitate from the cell fraction, or lane 8, the supernatant fraction, of GPI-PLC treated cells. *b*, N-glycanase treatment of FC-1 immunoprecipitates. Lane 1, CD16-transfected COS cells, lane 2, nylon wool nonadherent PBMCs without N-glycanase treatment. Lane 3, CD16-transfected COS cells, lane 4, nylon-wool-nonadherent PBMCs with N-glycanase treatment.
Methods. Surface labelling and immunoprecipitations were carried out as in ref. 19. Digestions with GPI-PLC (from *B. thuringiensis*) were carried out as in ref. 19 in 270-mM sucrose 10% PBS at 37 °C for 1 h using 40 μg ml^{-1} of enzyme. N-glycanase digestion of FC-1 immunoprecipitates was carried out according to the vendor's recommendations (Genzyme) in 0.17% sodium dodecyl sulphate (SDS), 0.2 M sodium phosphate *p*H 8.6, 10 mM phenanthroline hydrate, 1.25% NP-40 and 10 units per ml of enzyme for 18 h at 37 °C. Relative molecular mass (K) markers are given on the left of each panel.

K562 (chronic myelogenous leukaemia), HEL (erythroleukaemia), U937 (promonocytic leukaemia), HL60 (promyelocytic leukaemia), HUT102, Molt 4 and Jurkat (T-cell leukaemias), Raji and Daudi (Burkitt lymphomas), JY (lymphoblastoid cell line) and HepG2 (hepatoblastoma). Blot hybridization of genomic DNA from placenta revealed a pattern consistent with a single-copy gene (Fig. 2*b*).

The cDNA sequence spans 888 nucleotides (Fig. 3*a*), and encodes a predicted peptide of 233 residues. The first 18 residues have the typical features of a secretory signal sequence[28]. Following the signal peptide are two immunoglobulin-related segments (residues 25-110 and 111-196), each showing significant similarity to members of the immunoglobulin superfamily, especially to members of the constant-region C2 set[29]. Quantitative sequence comparisons using the ALIGN program of the National Biomedical Research Foundation protein identification resource[30], showed that CD16 is most closely related to the α-form of the murine Fcγ2b/1 receptor[31] followed by the human CD32 (refs 7-9) and murine β-forms[31,32] (Fig. 3*c*).

The polypeptide sequence ends with a short hydrophobic domain (residues 200-220) followed by four hydrophilic residues, only one of which is charged. A similar structure and hydropathicity profile (Fig. 3*b*) is shared by membrane proteins bearing glycosyl phosphatidylinositol phospholipid (GPI-PL)-linked carboxyl termini[33]. To determine whether the pCD16 encoded receptor was GPI-linked, we surface labelled COS cells expressing CD16 with ^{125}I. These were then digested with glycosyl phosphatidylinositol-specific phospholipase C (GPI-PLC), centrifuged and the resulting cell and supernatant fractions immunoprecipitated with mAB FC-1. All the labelled CD16 was found in the supernatant fraction (Fig. 4*b*). When nylon-wool non-adherent PBMCs (T cells and 'null' cells) were similarly treated, ~50% of the CD16 was released into the soluble fraction (Fig. 4*a*). Previous work has shown that a substantial proportion (up to 90%) of some phospholipid-linked proteins is refractory to GPI-PLC treatment[33], raising the possibility that all lymphocyte CD16 is linked to GPI.

To further investigate the relationship between the pCD16-encoded and lymphoid CD16 polypeptides, immunoprecipitates from transfected COS cells and nylon wool non-adherent PBMCs were digested with N-glycanase to remove all N-linked carbohydrate. The sizes of the resulting polypeptide backbones from both the COS cell and lymphoid molecules were identical (relative molecular mass (M_r) ≈ 26,000 (26 K); Fig. 4*b*) and agreed well with the relative molecular mass predicted from the peptide sequence (25K). Thus, the GPI-PL-linked form of CD16 appears to be the predominant, if not exclusive, form of CD16 on lymphoid cells *in vivo*. S1 analysis of poly(A)$^+$ RNA from LAK cells and short-term bulk cultures of PBMCs enriched for CD16$^+$ cells[27] confirmed this. The pCD16 cDNA insert protected only a single species of the expected size (≈900 base pairs (bp)), excluding the presence of RNAs encoding alternative cytoplasmic domains (Fig. 2*c*). Thus if conventionally anchored CD16 exists in lymphocytes, it must use the short hydrophilic terminus encoded in the cDNA.

CD16 appears to be the human homologue of the α-form of the murine Fcγ2b/1 receptor, which bears conventional anchoring sequences. The similarity between CD16 and murine α-receptors extends through the entire transmembrane domain (Fig. 3*c*) suggesting that removal of anchoring sequences may be sufficient to initiate phospholipid attachment. An intriguing aspect of the IgG Fc receptors so far isolated is the contrast between the conservation of their extracellular domains and the diversity of their cytoplasmic domains[8,9,32,33]. The diversity of the cytoplasmic region of the low-affinity receptors may reflect a signal transduction mechanism which proceeds through an extracellular intermediate, or 'receptor for Fc receptor'.

We thank Andrew J. McMichael and David Mason for providing us with the myeloid panel from the Third Leukocyte Typing Workshop, Janet Allen and Andrew Peterson for discussions, Sidney Udenfriend (Roche Institute) for a gift of *B. bacillus thuringiensis* GPI-PLC, Clark Anderson for a gift of human IgG3, Ivan Stamenkovic for B-cell RNAs and Alan Williams for providing details of the V, C1 and C2 set sequences[29]. This work was supported by a grant from Hoechst AG.

Received 7 April; accepted 5 May 1988.

1. Anderson, C. L. & Looney, R. J. *Immunol. Today* **7**, 64-66 (1986).
2. Burton, D. R. *Molec. Immun.* **22**, 161-206 (1985).
3. Anderson, C. L. *J. exp. Med.* **156**, 1794-1805 (1982).
4. Anderson, C. L. *et al. J. biol. Chem.* **261**, 12856-12864 (1986).
5. Dougherty, G. J., Selvendran, Y., Murdoch, S., Palmer, D. G. & Hogg, N. *Eur. J. Immun.* **17**, 1453-1459 (1987).
6. Looney, R. J., Abraham, G. N. & Anderson, C. L. *J. Immun.* **136**, 1641-1647 (1986).
7. Stuart, S. *et al. J. exp. Med.* **166**, 1668-1684 (1987).
8. Hibbs, M. L., Bonadonna, L., Scott, B. M., McKenzie, I. F. C. & Hogarth, P. M. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
9. Stengelin, S., Stamenkovic, & Seed, B. *EMBO J.* (in the press).
10. Fleit, H. B., Wright, S. D. & Unkeless, J. C. *Proc. natn. Acad. Sci. U.S.A.* **79**, 3275-3279 (1982).
11. Kulczycki, A. *J. Immun.* **133**, 849-854 (1984).
12. Perussia, B. & Trinchieri, G. *J. Immun.* **132**, 1410-1415 (1984).
13. Perussia, B. *et al. J. Immun.* **133**, 180-189 (1984).
14. Perussia, B., Starr, S., Abraham, S., Fanning, V. & Trinchieri, G. *J. Immun.* **130**, 2133-2137 (1983).
15. Perussia, B. *et al. J. Immun.* **130**, 2142-2148 (1983).
16. Malavassi, F. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 2443-2447 (1986).
17. Anegon, I., Cuturi, M. C., Trinchieri, G. & Perussia, B. *J. exp. Med.* **167**, 452-472 (1988).
18. Zarcone, D. *et al. Blood* **69**, 1725-1736 (1987).
19. Seed, B. *Nature* **329**, 840-842 (1987).
20. McMichael, A. J. *et al. Leukocyte Typing III. White Cell Differentiation Antigens* (Oxford University Press, Oxford, 1987).
21. Horovitz, A. & Levitzki, A. *Proc. natn. Acad. Sci. U.S.A.* **84**, 6654-6658 (1987).
22. Cheng, Y. & Prusoff, W. M. *Biochem. Pharmac.* **22**, 3099-3108 (1973).
23. Kipps, T. J., Parham, P., Punt, J. & Herzenberg L. A. *J. exp. Med.* **161**, 1-17 (1985).
24. Anasetti, C. *et al. J. Immun.* **138**, 2979-2981 (1987).
25. Spiegelberg, H. L., Perlmann, H. & Perlmann, P. *J. Immun.* **116**, 1464-1471 (1976).
26. Gergely, J., Sasmay, G., Rozsnyay, Z., Stanworth, D. R. & Klein, E. *Molec. Immun.* **23**, 1203-1209 (1986).
27. Perussia, B. *et al. Nat. Immun. Cell Growth Reg.* **6**, 171-188 (1987).
28. von Heijne, G. *Nucleic Acids Res.* **14**, 4683-4690 (1986).
29. Williams, A. F. *Immunol. Today* **8**, 298-303 (1987).
30. Dayhoff, M. O., Barker, W. C. & Hunt, L. T. *Meth. Enzym.* **91**, 524-545 (1983).
31. Revetch, J. V. *et al. Science* **234**, 718-725 (1986).
32. Lewis, V. A., Koch., T., Plutner, H. & Mellman, I. *Nature* **324**, 372-375 (1986).
33. Ferguson, M. A. J. & Williams, A. F. *A. Rev. Biochem.* (in the press).
34. Maniatis, T., Fritsch, E. F. & Sambrook, J. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, New York, 1983).
35. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463-5467 (1977).
36. Kyte, J. & Doolittle, R. *J. molec. Biol.* **157**, 105-113 (1982).

Stimulation of B-cell progenitors by cloned murine interleukin-7

Anthony E. Namen, Stephen Lupton, Kathryn Hjerrild, Janis Wignall, Diane Y. Mochizuki, Ann Schmierer, Bruce Mosley, Carl J. March, David Urdal, Steven Gillis, David Cosman & Raymond G. Goodwin

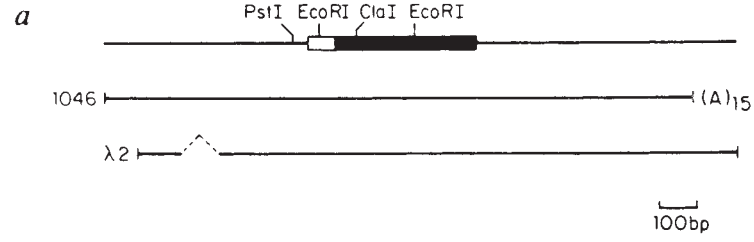
Immunex Corporation, 51 University Street, Seattle, Washington 98101, USA

The events involved in the commitment and development of lymphoid lineage cells are poorly understood. We have used a recently described long-term culture system[1] to establish a bioassay that can detect a novel growth factor capable of stimulating the proliferation of lymphoid progenitors[2]. Using direct expression in mammalian cells we have isolated a complementary DNA clone encoding this novel haematopoietic growth factor, designated interleukin-7.

A stromal cell line derived from murine bone marrow (IXN/A6)[2] was used to prepare a cDNA library, which was screened for biologically active interleukin-7 (IL-7) following expression in COS-7 cells (Fig. 1). Approximately 720,000 recombinants were screened before a pool yielding IL-7 biological activity was identified. The positive pool (244B) was subdivided until a single clone encoding IL-7 (clone 1046) was identified. Table 1*a* demonstrates the anticipated increase in bioactivity recovered from the COS-7 cell supernatants during the purification of the 1046 cDNA. There was a 2,000-fold enrichment of bioactivity as the cDNA pool size decreased to the single clone, 1046.

The nucleotide sequence of clone 1046 was determined and is shown in Fig. 1. The insert is 1,607 base pairs long, containing only one substantial open reading frame of 462 base pairs (bp). This was preceded by a 548-bp 5′ non-coding region containing several potential translation initiation sites. The open reading frame was followed by a 3′ non-coding region consisting of 579 bp, containing a consensus polyadenylation signal and terminating with a string of 15 adenine residues.

IL-7 purified from the IXN/A6 cell line[2] was subjected to N-terminal amino-acid sequence analysis. The results were compared to the predicted amino-acid sequence from clone 1046, allowing us to establish the N-terminus of the mature protein. IL-7 contains a 25 amino-acid leader sequence, followed by 129 amino acids, predicting a relative molecular mass (M_r) of ~14,900 (14.9K). The sequence contains two potential *N*-linked



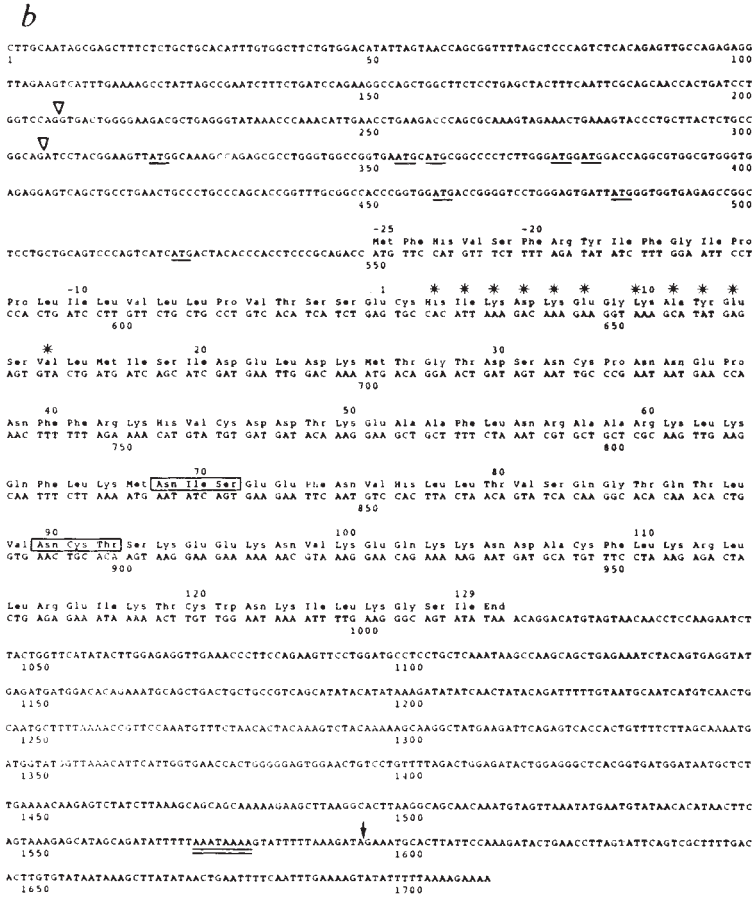


Fig. 1 *a*, Schematic representation of murine IL-7 cDNA clones. Dotted lines illustrate the sequences lacking in λ2. The open box illustrates the leader sequence of IL-7 and the solid box the region coding for mature IL-7. *b*, Nucleotide sequence and deduced amino acid sequence of clone 1046. Numbers above the sequences refer to amino acid positions and those below the sequence to the corresponding nucleotides. Inverted triangles represent the ends of the deletion in λ2. Additional initiation codons in the 5′ non-coding region are underlined. Potential *N*-linked glycosylation sites are boxed. Double underline, presumptive polyadenylation recognition site; vertical arrow, end of clone 1046. Asterisks, amino acids matching those determined by sequencing of the native protein.

Methods. Cytoplasmic mRNA from the IXN/A6 cell line was isolated as described by Pennica *et al.*[8]. Polyadenylated RNA was isolated and used to prepare double-stranded cDNA. Size-selected cDNAs were ligated into pDC201 vector DNA and transformed into *Escherichia coli* strain DH5α. Transformants were plated to provide ~1,000 colonies per plate. The resultant colonies were harvested and each pool used to prepare plasmid DNA for transfection into COS-7 cells as described[3,9], and the supernatants analysed for the presence of biologically active IL-7. The positive pool was subcloned and assayed until a single clone(number 1046) was isolated which encoded biologically active IL-7. The cDNA inserts were sequenced as described[10,11] and are shown in Fig. 1*b*. Amino acid sequencing was performed as described previously[12].

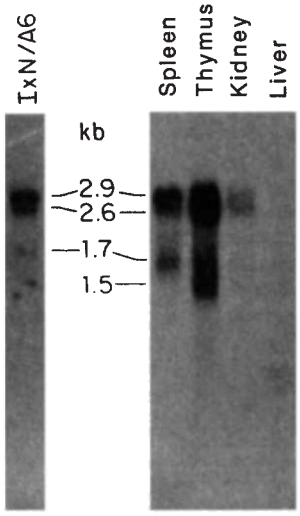


Fig. 2. Northern blot analysis of IL-7 mRNA prepared from various BALB/c tissues and the cell line IXN/A6. Total cellular RNA from the various murine tissues was isolated as previously described[14]. For Northern analysis, 10 μg of poly(A^+) RNA was fractionated on a 1.1% agarose-formaldehyde gel and blotted onto Hybond (Amersham) as recommended by the manufacturer. The blots were probed for IL-7 mRNA using an antisense ^{32}P-labelled transcript prepared by SP-6 RNA polymerase transcription of a pGEM1 plasmid containing the cDNA from clone 1046. The blots were hybridized and washed as described[14] and exposed to X-ray film at −70 °C with intensifying screens. The sizes of the IL-7 mRNAs were estimated by comparison with RNA standards obtained from Bethesda Research Laboratories.

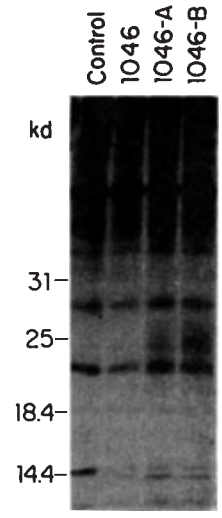


Fig. 3 Analysis of ^{35}S-labelled secreted IL-7 following expression in COS-7 cells of various cDNA constructs. Clone 1046-A has seven of the eight potential upstream initiation sites deleted, and clone 1046-B all eight of these sites deleted. The control lane is a COS-7 cell supernatant obtained following transfection with pDC201 which contained no inserted cDNA. The expressed IL-7 migrates at ~25K.

Methods. The vector pDC201 was designed to express inserted cDNA sequences when transfected into mammalian cells. The plasmid contains sequences from SV40 (the origin of DNA replication, the enhancer, early and late promoters and early polyadenylation signal) and adenovirus-2 (the major late promoter, tripartite leader and virus associated RNA genes as described elsewhere)[15]. Clone 1046A, which lacks the first 509 bp of clone 1046, was constructed by digestion of the isolated cDNA with *Pst*I and insertion of the subsequently blunted fragment into the *Sma*I site of pDC201. The various fragments used to construct clone 1046B were initially assembled in the GEMBL vector cut with *Bam*HI and *Hin*cII. Inserted into this vector were the 100 bp *Eco*RI/*Cla*I and the 400 bp *Cla*I/*Pvu*II fragments derived from clone 1046 and the synthetic oligonucleotides

GATCCCCACCATGTTCCATGTTTCTTTTAGATATATCTTTGG

GGGTGGTACAAGGTACAAAGAAAATCTATATAGAAACCTTAA

These oligonucleotides regenerate the translation initiation site (underlined) but eliminate all upstream initiation sites. The assembled construct was then removed from the GEMBL vector by digestion with *Sma*I and *Hin*dIII and the ends of the fragment blunted using T4-DNA polymerase. This fragment was then cloned into the *Sma*I site of pDC201. COS-7 cells were transfected as described previously[9] with the appropriate vector DNA three days prior to collection of media for IL-7 bioassay or labelling. For labeling, each plate of cells was washed twice with phosphate buffered saline (PBS) and incubated at 37 °C for 6 h in 2.5 ml MEM without methionine and cysteine but containing 100 μCi ^{35}S-methionine and ^{35}S-cysteine (Translabel, ICN Radiochemicals). The media containing labelled secreted proteins was removed and centrifuged, SDS sample buffer was added, and the samples were electrophoresed on an 18% SDS-polyacrylamide gel. The gel was treated with ENHANCE (NEN), washed, dried under vacuum, and exposed to X-ray film (Kodak XAR-5) for 3 days.

glycosylation sites at amino acids 69 and 90 (Fig. 1), the glycosylation of which probably accounts for the migration of the native protein at $M_r = 25K$[2]. There are six cysteine residues within the coding region of IL-7, some of which are likely to be involved in intramolecular disulphide bonds as treatment of IL-7 with 2-mercaptoethanol resulted in a loss of bioactivity (data not shown).

An additional cDNA clone (λ2) was isolated from an IXN/A6 library in λgt10 by hybridization with the insert from clone 1046. Compared with 1046, clone λ2 lacked the sequences from nucleotide number 208 to number 305 in the 5′ non-coding region. The sequences at the divergence points resemble consensus splice donor/acceptor sites, suggesting that the nucleotides were removed during the processing of the primary transcript. Clone λ2 also lacks the 15 adenine residues found at the 3′ end of the 1046 insert, but does contain additional bases extending beyond the 3′ end of 1046. The coding sequence of λ2 is, however, identical to that of clone 1046. Thus, it appears that IL-7 gene transcripts can undergo differential splicing and polyadenylation (see below).

Polyadenylated messenger RNA was prepared from several murine tissues, including thymus, spleen, kidney, liver and the IXN/A6 cell line and analysed by Northern blotting using the IL-7 cDNA as a probe. The patterns of RNA species hybridizing to the probe were complex and differed significantly depending on the tissue examined (Fig. 2). Transcripts were present at varying levels corresponding to 2.9, 2.6, 1.7 and 1.5 kilobases. The multiple species of mRNA observed may be the result of the use of alternative splicing and/or polyadenylation sites, as observed in the comparison of clones 1046 and λ2. We have not, however, been able to detect biologically active IL-7 in supernatants prepared from spleen or thymus. These transcripts may represent alternative splicing events leading to loss of biological activity as observed with the human IL-2 receptor[3] and G-CSF[4]. This will be examined by the isolation of cDNAs using mRNA derived from the spleen or thymus tissue.

Clone 1046 contains a large 5′ noncoding region with eight potential translational initiation codons upstream from the presumptive authentic initiation codon. We suspected that these potential upstream sites might decrease the translational efficiency of the IL-7 mRNA, as has been observed elsewhere[6]. Modifications of clone 1046 were prepared, expressed in COS-7 cells, and the ^{35}S-labelled secreted proteins analysed by electrophoresis and autoradiography (Fig. 3). A diffuse band could be seen in the 1046-A lane at 25K (the molecular size of native IL-7) and a more prominent band of the same size in the 1046-B lane. The amount of active IL-7 recovered correlated well with the intensity of the 25K band (Table 1*b*). Removal of the 5′ non-coding region improves expression of the cDNA in the COS-7 cell system and suggests that this region may be involved in the regulation of expression of IL-7 *in vivo*. The regulation of expression may aso occur by alternative splicing and polyadenylation as evidenced by the complex mRNA patterns observed by Northern analysis and the nature of the clones 1046 and λ2.

Table 1 Biological activity

a

Sample	IXN/A6	Control pDC201	Primary Pool 244B	Secondary Pool B.19	Clone cDNA 1046
Units ml^{-1}	20	0	5.5	58	11,334

b Effect of 5′ untranslated region on expression of IL-7 in COS-7

Sample	1046	1046-A	1046-B
Units ml^{-1}	11,334	38,399	111,203

c Response of cells to native and recombinant IL-7

LTBMC population	Native IL-7 maximum c.p.m. ($\times 10^{-3}$)	Recombinant IL-7 maximum c.p.m. ($\times 10^{-3}$)	Control
Nonadherent	44±4	43±4	2±0.5
$B220^+$ Nonadherent	41±6	56±8	2±0.5
$B220^-$ Nonadherent	51±5	72±7	2±0.6

a, The IXN/A6 sample is a serum-free supernatant derived from the IXN/A6 cell line as described[2]. Sample pDC201 is a COS-7 cell supernatant derived from transfection with the pDC201 plasmid without a DNA insert. Sample 244B is the supernatant from a transfected pool of 1,000 transformants. B.19 refers to the supernatant from a pool of 200 colonies derived from 244B. Clone 1046 is the purified IL-7 expressing clone. *b*, Clone 1046 and the derivatives 1046 A and B are described in Fig. 3. *c*, The nonadherent long term bone marrow cells (LTBMC) were fractionated with anti-B220 (ref. 7) coated magnetic beads as recommended by the manufacturers (Dynal, Great Neck, NY). The bioassay was performed as descrbed earlier[2]. Recombinant IL-7 is a COS-7 supernatant collected after transfection with clone 1046-B. Native IL-7 is purified from the supernatants of the IXN/A6 cell line as described[2]. The control was supernatants from COS-7 cells transfected with the pDC201 vector alone.

A clonal source of IL-7, a specific and senstive bioassay, and a high level cDNA expression vector have enabled us to isolate the cDNA encoding IL-7. Both recombinant IL-7 purified from COS-7 cell supernatants and native IL-7 had a similar specific activity of 4×10^6 units per μg. The repertoire of cells capable of responding to IL-7 is presently unknown. We have fractionated the non-adherent cells from our long term cultures into $B220^-$ (pro-B cells) and $B220^+$ (pre-B cells) and determined their ability to respond to native and recombinant IL-7 (Table 1*c*). Both cell populations responded equally well to natural and recombinant IL-7. Given the tissue distribution of IL-7 transcripts (Fig. 2) and the observation that both B220 positive and negative populations can respond to IL-7 (Table 1*c*), it is likely that IL-7 acts on a number of cell types. A clear assessment of the IL-7 responsive cellular populations and the array of cells displaying IL-7 receptors can now be determined using purified recombinant IL-7.

We thank Linda Troup for her asistance in the preparation of this manuscript, Jeff Hesselberg and Steve Gimpel for technical asisstance.

Received 23 February; accepted 21 April 1988.

1. Whitlock, C. A., Robertson, D. & Witte, D. N. *J. Immunol. Meth.* **67**, 353-369 (1984).
2. Namen, A. *et al. J. exp. Med.* **167**, 988-1002 (1988).
3. Cosman, D. *et al. Nature* **312**, 768-771 (1984).
4. Nagata, S. *et al. Nature* **319**, 415-418 (1986).
5. McCutchan, J. H. & Pagano, J. S. *J. natn. Cancer Inst.* **41**, 351-357 (1968).
6. Kozak, M. *Nucl. Acids Res.* **12**, 3873-3893 (1984).
7. Kincade, P. W., Lee, G., Watanabe, T., Sun, L. & Scheid, M. P. *J. Immunol.* **127**, 2262-2268 (1981).
8. Pennica, D. *et al. Nature* **301**, 214-221 (1983).
9. Luthman, H. & Magnusson, G. *Nucleic Acid Res.* **11**, 1295-1308 (1983).
10. Hattori, M. & Sakaki, Y. *Analyt. Biochem.* **152**, 232-238 (1986).
11. Norrander, J., Kempe, T. & Messing, J. *Gene* **26**, 101-106 (1983).
12. March, C. J. *et al. Nature* **315**, 641-646 (1985).
13. Devereux, J., Haeberli, P. & Smithies, D. *Nucleic Acid. Res.* **12**, 387-395 (1984).
14. Watson, J. D. *et al. J. Immun.* **139**, 123-129 (1987).
15. Sims, J. *et al. Science* (submitted).

The phylogenetic history of immunodeficiency viruses

T. F. Smith*, A. Srinivasan†, G. Schochetman†, M. Marcus* & G. Myers‡

* MBCRR Dana-Farber Cancer Institute, Harvard University, Boston, Massachusetts 02115, USA
† Center for Infectious Diseases, Centers for Disease Control, Atlanta, Georgia 30333, USA
‡ T-Divison, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

Knowledge of the phylogenetic history of the human immunodeficiency viruses (HIV-1 and HIV-2) is important for our understanding of the epidemiology of AIDS, the disease caused by these viruses. Reconstruction of the evolutionary tree is hampered, however, by two problems. One is the high variation in nucleotide sequence between the known HIV isolates[1,2] which can create formidable difficulties in identifying homologous genomic sites that may be used in a molecular phylogenetic reconstruction. Another impediment has been the lack of unequivocal time calibration points: there is only a sparse 'fossil record' for HIV and limited historical epidemiological data. We have largely overcome these difficulties by: (1) a thorough optimal-sequence alignment analysis; (2) the inclusion of sequences of an early (1976) HIV-1 isolate,[3] a recent (1986) HIV-2 isolate[4] and two simian immunodeficiency viruses (SIV)[5,6] along with five other HIV-1 isolates[7,8-14]; and (3) the reconstruction of a minimum-length evolutionary tree based on the envelope-gene variable positions. We conclude that HIV-1 may have evolved from its common ancestor with HIV-2 as recently as 40 years ago.

Calculation of the tree with the minimum mutation distances for a set of sequences, based on the information from every postion that varies between the sequences, is referred to in numerical taxonomy as the method of 'maximum parsimony'[15-21]. This method depends on the pairing of all homologous sequence positions as accurately as possible, which accordingly requires accurate and optimal alignment of all the sequences. The envelope genes of the known HIV and SIV isolates provide sufficient variation to make discrimination of the branch order of the tree robust—that is, insensitive to minor alignment errors or alternatives—while still providing sufficient sequence conservation to allow nearly unambiguous sequence alignment.

The envelope sequences contain three taxonomically useful sites with conserved structure and function: the conserved cysteines; the putative glycosylation sites; and the protein cleavage sites. Using a standard dynamic programming algorithm[22,23], initial alignments between all sequence pairs were obtained based only on the relative spacing of the cysteine residues. These alignments were then optimized locally[23] over the regions between the aligned cysteine residues to minimize non-conservative amino-acid replacements. Finally, the remaining ambiguous regions were optimally aligned by minimizing the number of implied nucleotide replacements[23]. The resulting pairwise alignments appear to conserve the majority of the glycosylation site positions (Fig. 1) and show a strong correlation between the aligned hydrophobicity profiles.

These optimal alignments between pairs of sequences were then combined to construct the overall alignment of all nine viral sequences. This was done under the constraint of maintaining the pairwise alignment of the conserved cysteines and the putative glycosylation sites and ambiguities were adjusted to minimize the number of nucleotide replacements. Figure 1 depicts the resulting alignment of the envelope protein sequence. While this procedure cannot guarantee perfect alignment of all homologous sites, it will err in the direction of underestimating

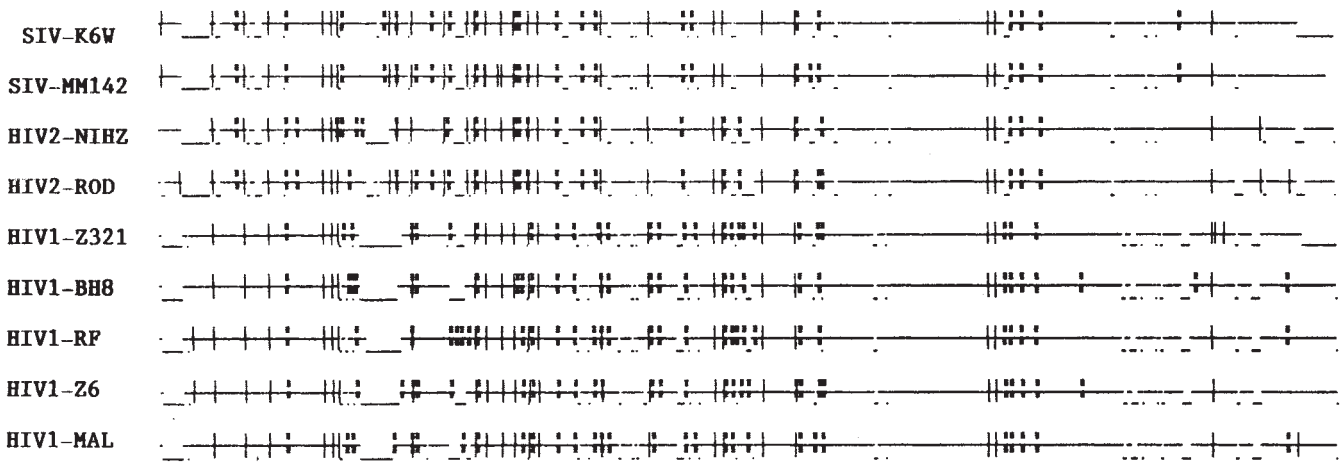


Fig. 1 Schematic depiction of the alignment of nine AIDS viral envelope protein sequences. Each sequence is represented by a horizontal line that is broken and underlined at positions where optimal alignment required gaps. Vertical lines mark the locations of cysteine residues. Vertical block-pairs mark potential glycosylation sites. The virus isolates are: SIV_{K6W78}, an originally misidentified HIV-2 1984 isolate, probably a Macaque contaminate[6]; SIV_{MM142}, a Macaque 1983 isolate from the New England Primate Center[5]; HIV-2_{NIHZ}, a 1986 isolate from an African patient[4,8]; HIV-2_{ROD}, a 1984 isolate from a Guinea-Bissau patient[9]; HIV-1_{Z321}, isolated from a stored 1976 Zairian blood sample[24] isolated at the Centers for Disease Control CDC in 1987; HIV-1_{BH8}, a 1983 isolate from pooled blood of New York City patients[7]; HIV-1_{RF}, a US isolate from an AIDS patient of Haitian origin[11]; HIV-1_{Z6}, isolated at the CDC in 1983 from a Zairian blood sample[13]; HIV-1_{MAL}, a 1984 isolate from a 7-year-old Zairian, probably infected by a 1980 blood transfusion[14].

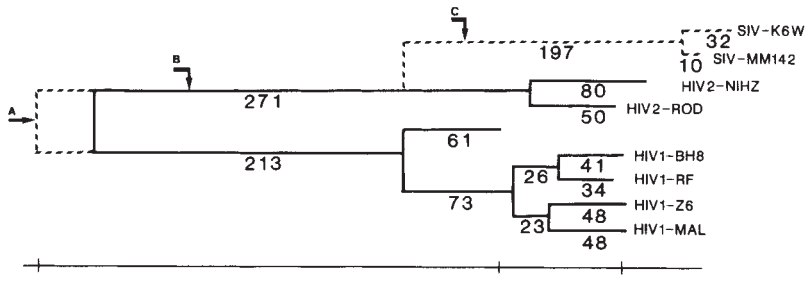


Fig. 2 Minimum-length tree constructed from variation in the envelope protein third codon position using PAUP[17] with the global branch-swapping option, MULPARB. The virus identification is the same as in Fig. 1. The horizontal lines have lengths proportional to the minimum number of nucleotide substitutions required to generate the observed variation. The length of the vertical lines is for clarity only. The solid-line HIV subtree is the most minimal of all possible branching patterns, derived from the ALLTREES option. The three arrows indicate possible alternate routes: A represents approximately equal HIV-1 and HIV-2 rates of nucleotide substitution; B represent a ratio of HIV-1 to HIV-2 rates of 3:2: represents a ratio of HIV-1 to combined HIV-2 and SIV-2 rate of 4:1. The dashed lines at proposed root (A) of the tree represent an approximate multiple-hit correction.

overall divergence in regions of minimal constraints, that is regions of maximum variability. The result is a conservative estimate of the sequence differences and therefore of the implied divergence times.

Assuming that homologous sites have been correctly aligned at the vast majority of sequence positions, one can obtain the number of first, second and third codon positions at which at least one point mutation has occurred (see Table 1). In codon position three, for example, over half of the codons not involved in a deletion/insertion event in any of the nine isolates have suffered at least one mutation since the divergence of HIV-1 and HIV-2. Most of the apparent nucleotide substitutions in codon position three are transitions. They may thus be considered 'neutral' mutations and could be expected to have accumulated at a rate proportional to the generation time, or average number of genome replications along any particular line of descent.

A minimum-length tree was generated from the variation in each separate codon position using the parsimony algorithm PAUP of Swofford[18,19]. All three positions generate the same branching pattern, albeit with slightly different relative branch lengths. Notably the variation at positions two and three generates the same tree from respectively the most and the least structurally constrained codon positions. To verify this approach, we also used a simple distance-linkage method and the method of Transversion parsimony[21]; all three methods gave an identical branching order.

The branching order depicted in Fig. 2 is a robust function of the current data and is compatible with the phylogenetic relationships previously generated from a larger set of envelope sequences from HIV-1 only[24]. While there is a second minimum-length tree compatible with the codon position two data, it differs only in that the two African sequences, HIV-1_{MAL} and HIV-1_{Z6} branch separately rather than together. In addition to generating trees using each codon position separately, the branching pattern was confirmed by performing the phylogenetic analyses on all possible data subsets containing only eight of the nine sequences, and on the first half versus the second half of the envelope gene. In each case the results were identical to those of Fig. 2. Finally, a phylogenetic reconstruction was carried out using the least reliable alignment information, the deletion positions. Even in this case one of the three minimum-length trees had the same branching order. Rather than rooting the tree at an equal distance from the HIV-1 and the HIV-2 cluster, it is possible to root it on the branch that connects the HIV-2 and SIV clusters. This would, however, require the distance of the last common ancestor of HIV-1, HIV-2 and SIV to differ from that shown in Fig. 2 by a factor of three to five and currently appears inappropriate without direct supporting data.

As there is limited information on the determinants of mutation and fixation rates in these viruses, the above calculations are based on the assumption that the rate of fixation has been similar throughout the recent history of these viruses. This amounts to accepting the 'uniform average rate' hypothesis, which has proven applicable to many studies[20]. This is not a constant clock assumption. It is rather the recognition that with sufficient time and number of positions the effects of the many factors influencing mutation and fixation rates (such as protein function, viral pathogenicity, generation time, host variation, rate of pandemic spread, and so on) often result in a nearly-uniform average rate. This result is directly related to the statistical law of large numbers.

The tree shown in Fig. 2 appears consistent with such a minimal assumption. All of the HIVs isolated between 1983 and 1985 are at nearly equal mutational distances from a hypothetical HIV-2/HIV-1 common ancestor. The major exception is the oldest known isolate, HIV-1_{Z231}, from 1976[3]. By assuming that the seven-year difference in isolation between the HIV-1_{Z321} and the other HIV-1 isolates is the main reason for the difference in branch length (from such a common HIV-1/HIV-2 ancestor) the tree branch lengths might be calibrated as a function of time. This depends on there being nothing atypical about the 1976 isolate.

Under the above assumption, the average difference between the 1976 isolate and the other HIV-1s is 11.6 ± 2.0 substitutions per thousand per year in codon position three. In comparison the isolation of the two HIV-2s differs by about $1\frac{1}{2}$ years[8,9], and the mutational distance difference from their last common ancestor is thirty substitutions in the variable position-three sites, resulting in about 21 ± 5 substitutions per thousand per year. The two values are within a factor of two, and are comparable to those for the haemagglutinin and non-structural protein (NS) genes of the influenza A virus (ref. 25 and W. Fitch, personal communication).

Using the HIV-1 substitution value of 11.6 gives a date of 1951, plus or minus 3 years, as the most recent equidistant HIV-1/HIV-2 divergence date. It must be emphasized that this estimate derives from a minimal-variation alignment and a minimal assumption with regard to the evolutionary dynamics involved. Therefore, 37 years appears to be a lower bound to the age of this disease. By any hypothesis, however, the onset of conspicuous diversification of these viruses correlates strongly with the known historical rise of the pandemic, although the epidemiological record may be insufficient with respect to the very earliest occurrences. It is clearly difficult to reconcile our results with the view that the observed variation has existed independently in the various lines of descent leading to these isolates for a long time[26]. This is particularly true if one examines the phylogenetic clustering of the sixteen known HIV-1 isolates where clusters correspond closely with the geographic regions of isolation[24].

Table 1 Phylogenetic variation per codon position in nine AIDS envelope sequences, aligned as in Fig. 1

Codon positions	One	Two	Three
Total aligned sites	964	964	964
Overlapped by a deletion*	238	238	237
Replaced at least once	433	374	601
Minimum required mutations†	563	493	905
Minimum length trees	774	674	1,209‡

* All positions overlapped by a deletion in any one of the nine sequences were not used in construction of the minimum length trees.

† The minimum required number of mutations (if every position was compatible with the same branching pattern) is the sum of the number of different nucleotides observed at each position in the aligned data set, minus one.

‡ Length of tree shown in Fig. 1.

The calibrated phylogenetic relationship presented here invites the interpretation that viral diversification has proceeded in step with the epidemics in Africa, Haiti and in the United States. The earliest serological datum for the virus stems from one possible case in Zaire in 1959[27]. Serological and clinical evidence reveals a marked increase in HIV infection in Africa in the mid-1970s (refs 1, 28) and the earliest evidence of HIV-1 in Haiti and the United States appears to be in the mid-1970s (refs 29, 34), with one intriguing possible exception in the United States in 1969. An isolated 1969 case of a teenager with the rare Kaposi's sarcoma, who died in a St Louis hospital, has now been shown to have had AIDS (M. Witte, M. Elvin-Lewis, R. Garry and A. Gottlieb, unpublished data).

There is neither sufficient sequence information nor sufficient samples of viruses at this time to assess the accuracy of a simple average linear relationship between substitution rate and time. No known correction would reduce the current estimate, however, and few seem likely to increase it by more than a factor of two—much less than an order of magnitude. The current correlation with the epidemiological data, not to mention with the modern urbanization of Africa following the Second World War, seems unlikely to be accidental[1]. Older sequence information for both human and simian immunodeficiency viruses would be especially illuminating, but the prospects of finding such 'fossil' data may be diminishing as the disease progresses.

We thank Genoveffa Franchini and Marvin Reitz, Jr. for making the $HIV2_{NIHZ}$ sequence available prior to publication, Arnout Eikeboom and Don Faulkner for programming, David Swofford for his PAUP, and the NIH/DRR for support of the MBCRR. We also thank James Lake (UCLA) for independently reanalysing the sequence trees. This work was supported by the AIDS Program of the National Institute of Allergy and Infectious Diseases and the US Department of Energy.

Received 20 November 1987; accepted 28 April 1988.

1. Piot, P. *et al. Science* **239**, 573-579 (1988).
2. Clavel, F. *AIDS* **1**, 135 (1987).
3. Getchell, J. *et al. J. infect. Dis.* **156**, 833-837 (1987).
4. Franchini, G. *et al. AIDS and Human Retrovirus Res.* **3**, 11-17 (1987).
5. Kanki, P. J. *et al. Science* **228**, 1199-1201 (1985).
6. Kanki, P. J., Alroy, J. & Essex, M. *Science* **230**, 951-954 (1985).
7. Ratner, L. *et al. Nature* **313**, 277-284 (1985).
8. Zagury, J. F. *et al. Proc. natn. Acad. Sci. U.S.A.* (in the press).
9. Clavel, F. *et al. Science* **233**, 343-346 (1986).
10. Clavel, F. *et al. Nature* **324**, 691-695 (1986).
11. Popovic, M. *et al. Science* **224**, 497-499 (1984).
12. Feorino, P. M. *et al. Science* **225**, 69-72 (1984).
13. Srinivasan, A. *et al. Gene* **52**, 71-82 (1987).
14. Alizon, M. *et al. Cell* **46**, 63-74 (1986).
15. Fitch, V. M. *Syst. Zool.* **20**, 406-416 (1971).
16. Fitrch, W. M. *Am. Nat.* **111**, 223-257 (1977).
17. Day, W. H. E. *J. theor. Biol.* **103**, 429-438 (1983).
18. Swofford, D. L. & Maddison, W. P. *Math. Biosci.* (in the press).
19. Swofford, D. L. *PAUP User's Guide* (Illinois Natural History Survey, Champaign, Illinois, 1985).
20. Lewin, R. *Science* **239**, 561-563 (1988).
21. Lake, J. *Nature* **331**, 184-186 (1988).
22. Waterman, M. S., Smith, T. F. & Katcher, H. L. *Nucleic Acids Res.* **12** (1984).
23. Waterman, M. S. & Smith, T. F. *Adv. appl. Math.* **2**, 482-489 (1981).
24. Myers, G., Josephs, S. S., Rabson, A. B., Smith, T. F. & Wong-Staal, F. (eds) *Human Retroviruses and AIDS, 1987* (AIDS Program of the National Institute of Allergy and Infectious Diseases and the US Department of Energy, Washington, DC, 1987).
25. Buonagurio, D. A. *et al. Science* **232**, 980-982 (1986).
26. Nahmias, A. J. *et al. Lancet* **i**, 1279-1280 (1986).
27. Quinn, T. C., Mann, J. M., Curran, J. W. & Piot, P. *Science* **234**, 955-963 (1986).
28. Curran, J. W. *et al. Science* **239**, 610-616 (1988).
29. Jaffe, H. W. *et al. Ann. intern. Med.* **103**, 210-214 (1985).
30. Jaffe, H. W. *et al. J. Am. med. Assoc.* **254**, 770-773 (1985).
31. Selik, R. M., Haverkos, H. W. & Curran, J. W. *Am. J. Med.* **76**, 493-500 (1984).
32. Darrow, W. W. *et al. Am. J. Publ. Health* **77**, 479-483 (1987).
33. Stevens, C. E. *et al. J. Am. med. Assoc.* **255**, 2167-2172 (1986).
34. Cossin, J. *Cell* **46**, 1-4 (1986).

Domain of *Ultrabithorax* expression in *Drosophila* visceral mesoderm from autoregulation and exclusion

Mariann Bienz & Gaby Tremml

Zoological Institute, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland

Domains of differential homeotic gene activity are formed at specific positions along the anteroposterior axis of the early *Drosophila* embryo[1]. Homeotic genes are required continuously throughout development[2], so that homeotic gene activity has to be maintained independently of the positional information provided in the early embryo. In the ectoderm, the domains of homeotic gene activity partially overlap[1], but we have found that in the visceral mesoderm at least three of these genes are expressed in adjacent and mutually exclusive domains. It has been proposed that stable, sharply demarcated domains of this type could be established if a homeotic gene product stimulated its own expression locally and inhibited the expression of other homeotic genes, which Meinhardt has termed autocatalysis and mutual exclusion respectively[3,4]. Furthermore, autocatalysis of this kind can in principle account for the maintenance of homeotic gene activity throughout development. We find that the unique domain of *Ultrabithorax* (*Ubx*) expression in the visceral mesoderm is dependent both on autocatalysis and on an exclusion mechanism: *Ubx* product is required for its own synthesis, whereas the product of the posteriorly adjacent gene *abdominal-A* represses *Ubx* expression.

We have previously demonstrated that the *Ubx* gene is expressed in parasegment 7 (ps7) of the visceral mesoderm in the *Drosophila* embryo[5]. This is in contrast to the domain of *Ubx* expression in the ectoderm (ps5–13)[6] and different from the genetically defined domain in the adult fly, where *Ubx* function is mostly required (ps5 and 6)[7]. Analysing the expression of various *Ubx*-β-galactosidase fusion genes (*Ubx*/β*gal*) in transformed embryos, we concluded that *Ubx* expression is dependent on different *cis*-acting sequences in the visceral mesoderm and in the ectoderm[5]. Thus, *Ubx* expression in the visceral mesoderm may be under the control of regulatory genes that are different from the set of genes controlling *Ubx* expression in the ectoderm[8–11]. Our genetic analysis aims to identify genes that regulate *Ubx* expression in the visceral mesoderm.

First, we asked whether *Ubx* product itself is required for its own expression. We constructed flies containing a *Ubx*/β*gal* fusion gene[5] (−3.1P1), as well as *Ubx* pseudo-point mutation[12,13] (Ubx^1, $Ubx^{9.22}$, Ubx^{195}), a *Ubx* deletion[12] (($Df(bxd^{100})$)), or a chromosome with null mutations in all three bithorax complex genes[14] (Ubx^{MX12} $abd\text{-}A^{M1}$ $Abd\text{-}B^{M8}$; referred to as *triple*). We then generated homozgous Ubx^- embryos by crossing −3.1P1 flies carrying the *triple* chromosome with −3.1P1 flies carrying one of the *Ubx* mutations. The results were essentially the same in all four crosses (Fig. 1): the strong β-galactosidase staining belt in the visceral mesoderm was absent in about one quarter of the embryos, the presumed Ubx^- homozygotes (Fig. 1*c*, *e*). Instead, these embryos show bulges in the region of the visceral mesoderm and/or the underlying midgut epithelium, where the blue belt is normally visible (Fig. 1*c*, *e*, *i*). In embryos homozygous for *Ubx* pseudo-point mutations (*inter se* crosses) we could see some residual β-galactosidase staining in addition to the bulges in ps7 of the visceral mesoderm (Fig. 1*e*). Nevertheless, none of the Ubx^- embryos shows the second of the three midgut constrictions (Fig. 1*d*, *f*, *j*) that are formed in normal stage-16 embryos[15]. In wild-type embryos, this second midgut constriction spatially coincides with the posterior limit of *Ubx* expression in the visceral mesoderm (Fig. 1*b*, *h*).

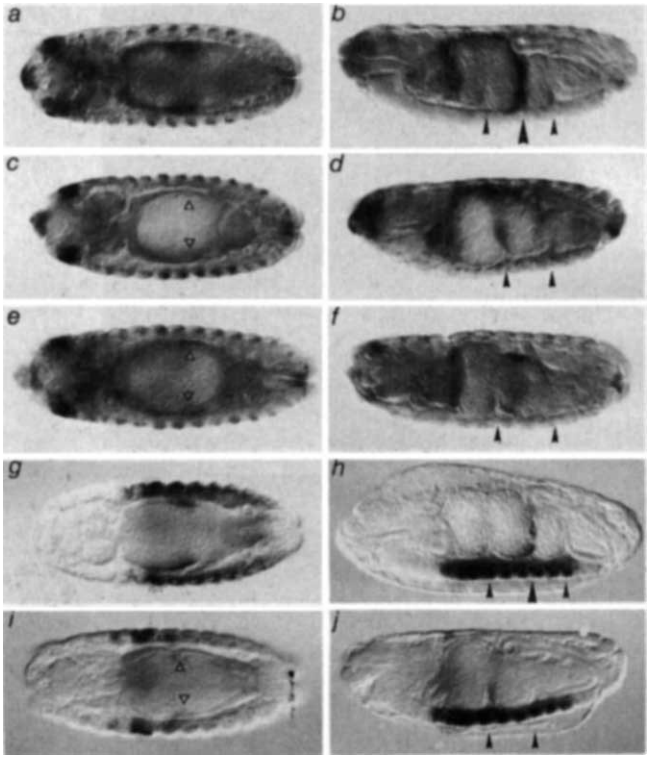


Fig. 1 Autocatalysis of *Ubx* expression. The expression of a *Ubx*/β*gal* fusion gene in wild type and in several homozygous *Ubx* mutant embryos was assayed by β-galactosidase staining at stage 15 (*a*, *c*, *e*) and 16 (*b*, *d*, *f*). Staining in ps7 of the visceral mesoderm is strong in the wild type (*a*), absent in a Ubx^1/*triple* embryo (*c*), and much reduced in a Ubx^1/Ubx^1 embryo (*e*); in both embryos that have no functional *Ubx* product (*c*, *e*) a bulge can be seen (triangle) where β-galactosidase staining in the wild type is visible. This bulge is typical for Ubx^- embryos, but is not observed in Ubx^+ embryos. Later, the posterior boundary of the β-galactosidase staining belt in the wild type (*b*) coincides with the second midgut constriction (large arrowhead) which is absent in all Ubx^- homozygotes (*d*, Ubx^1/*triple*; *f*, Ubx^{195}/Ubx^{195}), even if they show residual β-galactosidase staining (*f*). The first and the third midgut constrictions are indicated by small arrowheads. Antibody staining of endogenous Ubx protein in the wild type (*g*, *h*) and in antibody-positive homozygous *Ubx* mutants (*i*, *j*, $Ubx^{9.22}/Ubx^{9.22}$) confirms the results obtained by β-galactosidase staining. Antibody staining is generally weaker in these Ubx^- mutants, but only the visceral mesoderm pattern and not the ecoderm pattern of *Ubx* expression is noticeably affected by the mutation. Arrowheads are used as in *a*–*f*, and the staining procedures were as described previously[5].

Of all *Ubx* mutants we used, only $Ubx^{9.22}$ and Ubx^{195} mutant embryos accumulate Ubx protein[13] that can be detected with a Ubx antibody raised against the aminoterminal domain of Ubx protein[16]. Mutant Ubx protein in both cases lacks the homeodomain[13]. Antibody staining is generally weaker in these Ubx^- homozygotes, probably due to reduced stability of the mutant protein, but the pattern of *Ubx* expression appears unaltered in the ectoderm. In contrast, in the visceral mesoderm of the same embryos, Ubx antigen is barely detectable (*inter se* crosses) or completely absent (*triple* crosses). Instead, we observe the appearance of bulges and the lack of a second midgut constriction (Fig. 1*i* and *j*). Thus, the endogenous *Ubx* gene responds in the same way as the *Ubx*/β*gal* gene to the lack of functional Ubx protein. We conclude that *Ubx* expression in the visceral mesoderm depends on functional Ubx protein in an autocatalytic fashion. We do not know whether this autocatalysis acts at the level of initiation or maintenance of *Ubx* expression.

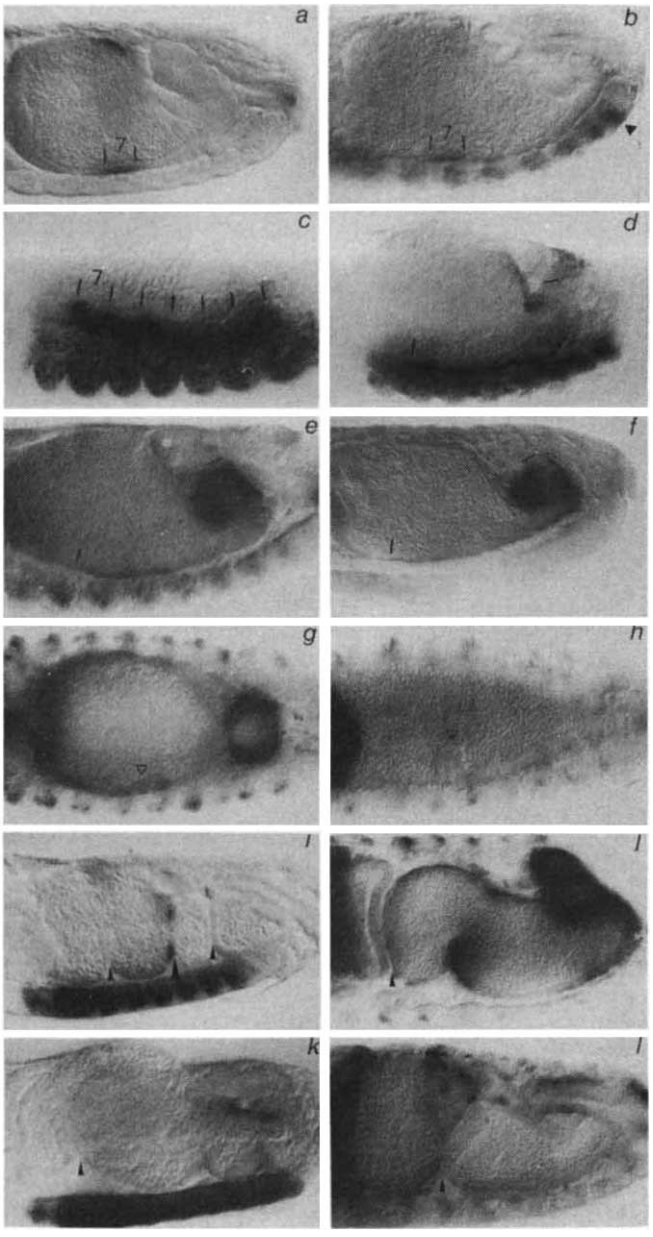


Fig. 2 Exclusion of *Ubx* expression by *abd-A*. Expression of a *Ubx/βgal* fusion gene (*a*, *e–h*, *j*, *l*) or of the endogenous *Ubx* gene (*b–d*, *i*, *k*) was assayed in stage 15 and 16 embryos as in Fig. 1. The *Ubx* expression domain in the visceral mesoderm is extended in *abd-A*$^-$ mutant embryos (*c*, *d*, *e*) by at least 5 parasegments (indicated in a stage 14 embryo in *c*; ps7 marked) to the end of the midgut (visible in *e*, abd-A^{M1}/*triple*). Bars indicate beginning and end of the normal and extended *Ubx* expression domains; it is virtually impossible to vizualise the whole domain in one focal plane. The extended *Ubx* expression domain is not altered in *abd-A*$^-$ *Abd-B*$^-$ mutants (*f*, *abd-A*^{M1} *Abd-B*^{M8}/*triple*). Similarly, the normal *Ubx* expression domain in the visceral mesoderm is not affected by mutation in *Abd-B* (*b*, *Abd-B*^{M1}/*Abd-B*^{M1}), although it is in the ectoderm[8] (strong staining in ps13 indicated by black triangle). But in the absence of functional *Ubx* and *abd-A* product (*g*, *Df109/triple*), *Abd-B* can activate the *Ubx/βgal* fusion gene ectopically in the visceral mesoderm of the posterior midgut (bars); the region of ps7–10 is unstained, but the bulge typical for *Ubx*$^-$ embryos is visible (triangle). The ectopic expression domain is lost in embryos that have no functional products from the bithorax complex (*h*, *DfP9/triple*; focused on bulge, triangle). Later, in the absence of functional *abd-A* product, only the first (small arrowhead) of the three midgut constrictions seen in the wild type (*i*) forms properly (*j*, *abd-A*^{M1}/*abd-A*^{M1}; *k*, *abd-A*^{M1} *Abd-B*^{M8}/*abd-A*^{M1} *Abd-B*^{M8}; *l*, *DfP9/triple*).

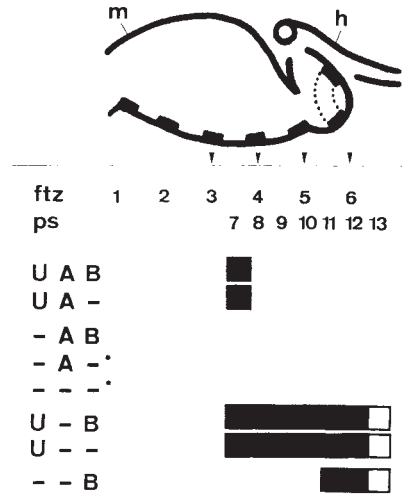


Fig. 3 Summary of the regulatory interactions. Six β-galactosidase stripes in a *ftz/βgal* transformant embryo[20] are indicated in a schematic drawing of a midgut (*m*) from a stage-15 embryo (*h*, hindgut, to show orientation). The two β-galactosidase stripes in the visceral mesoderm of the posterior midgut may represent the same (the sixth) *ftz* stripe. The normal (ps7), the extended (ps7–12) and the ectopic (ps11,12) *Ubx* expression domains are drawn next to the genetic constitutions that generate them (U, *Ubx*$^+$; A, *abd-A*$^+$; B, *Abd-B*$^+$; –, mutant). Open bars at the posterior end of the extended and the ectopic domains indicate that these may extend beyond ps12. Asterisks indicate the genetic constitutions of embryos that show weak ectopic β-galactosidase staining in the hindgut.

It appears that the mutant *Ubx*$^{9.22}$ or *Ubx*^{195} product is able to confer residual autocatalysis, whereas it is defective with respect to what might be the 'morphogenetic' function of *Ubx* in the visceral mesoderm, the induction of the second midgut constriction. The result implies that the 'morphogenetic' function, but perhaps not the 'regulatory' autocatalytic function, of *Ubx* is dependent on a fully functional Ubx protein, including the homeodomain.

We next asked whether *Ubx* expression in the visceral mesoderm is affected by the products of homeotic genes that are expressed in adjacent domains, the *Antennapedia* (*Antp*) and the *abdominal-A* (*abd-A*) genes. We stained embryos with antibody raised against Antp or with both Ubx and Antp antibody simultaneously and found that *Antp* is expressed in a single domain directly anterior to *Ubx*, probably in ps6. Posteriorly adjacent to *Ubx* in the visceral mesoderm, *abd-A* encodes protein expressed in a wide domain, starting in ps8 and extending to the posterior end of the midgut (F. Karch and W. Bender, personal communication).

We stained embryos homozygous for an *Antp* null allele[8] (*Antp*$^{NS+RC3}$) with Ubx antibody and found that *Ubx* expression was not affected by the *Antp* mutation. Likewise, the expression of the *Ubx/βgal* fusion gene is unaffected. We conclude that Antp function is not required for *Ubx* expression in the visceral mesoderm. The same conclusion was reached previously for *Ubx* expression in the ectoderm[8].

But mutation in the *abd-A* gene strikingly affects *Ubx* expression in the visceral mesoderm (Fig. 2). We set up crosses with flies carrying an *abd-A* null allele[14] (*abd-A*^{M1}) as described above, and stained offspring embryos for β-galactosidase activity or with Ubx antibody. We found that *abd-A*$^-$ homozygotes show derepression of *Ubx* in the ectoderm (Fig. 2*c*, *d*) as previously reported[8]. In the visceral mesoderm of these embryos, we observe an abnormal *Ubx* expression domain extending to the posterior end of the midgut (Fig. 2*c–e*). For reasons we do not understand, there is a slight difference between the

derepression of the fusion gene and of endogeneous *Ubx*: β-galactosidase staining appears weak anteriorly and strong posteriorly, whereas antibody staining is even throughout. But the extension of the derepressed domain is the same in both cases and most probably corresponds to the addition of the normal *Ubx* and *abd-A* expression domains. We conclude that functional *abd-A* product determines the posterior limit of normal *Ubx* expression in the visceral mesoderm. Interestingly, neither the second nor the third midgut constriction can form properly in *abd-A*$^-$ mutant embryos (Fig. 2*j*-*l*), suggesting that the formation of these constrictions depends on adjacent domains of differential homeotic gene activity in the visceral mesoderm.

The *Abdominal-B* (*Abd-B*) product functions to down-regulate *Ubx* expression in the posterior ectoderm[8]. In contrast, we find that an *Abd-B* null mutation[14] (*Abd-B*^{M1}) does not affect *Ubx* expression in the visceral mesoderm, although the effect of this mutation is clearly apparent in ps13 of the ectoderm (Fig. 2*b*). We find unexpectedly that, in the absence of functional *abd-A* product, *Abd-B* can activate the fusion gene ectopically: β-galactosidase staining appears in the visceral mesoderm of the posterior midgut in *Ubx*$^-$ *abd-A*$^-$ double mutant embryos (Fig. 2*g*). This ectopic β-galactosidase staining is abolished in triple mutants (*Ubx*$^-$ *abd-A*$^-$ *Abd-B*$^-$) and is therefore dependent on functional *Abd-B* product (Fig. 2*h*). The ectopic β-galactosidase staining domain probably corresponds to the *Abd-B* expression domain, implying that *abd-A* and *Abd-B* expression coincides in the posterior visceral mesoderm. It appears that *Abd-B* does not have any repressor function in this germ layer.

We have shown that, in the absence of functional *Ubx* product, the *Ubx* gene cannot be expressed in ps7, whereas, in the absence of functional *abd-A* product, *Ubx* expression cannot be repressed posterior to ps7 in the visceral mesoderm (Fig. 3). The targets for *Ubx* autocatalysis and *abd-A* represssion are evidently present on the *Ubx*/*βgal* fusion gene, and it is possible that these regulatory mechanisms are mediated by direct DNA binding[17] of the Ubx and abd-A homeoproteins to *Ubx* promoter sequences. It appears that in the visceral mesoderm autocatalysis and exclusion[2,3] are important for the establishment and/or the maintenance of adjacent domains of homeotic gene expression. Whether these mechanisms also act in the ectoderm where expression domains of different homeotic genes partly overlap[1] is unclear. Indication for autocatalysis in the ectoderm has been obtained for *fushi tarazu* expression[18], although in this case there is no evidence for exclusion by the adjacent gene *even-skipped*[19].

It is puzzling that only the posteriorly adjacent gene *abd-A*, but not the anteriorly adjacent gene *Antp*, excludes *Ubx* expression in the visceral mesoderm. It is possible that the anterior limit of the *Ubx* expression domain in this germ layer is determined by an additional regulatory protein that provides a positional value with respect to the antero-posterior axis. This hypothetical regulator may be present at low concentration anteriorly and at high concentration posteriorly and allow *Ubx* expression only above a certain threshold concentration. The posterior limit of *Ubx* expression could be determined by the exclusion function of the *abd-A* gene.

We thank Gines Morata for providing mutant flies, Rob White and Danny Brower for antibody, Hans Meinhardt and Hugh Pelham for critical discussion. This work was supported by the Swiss National Science Foundation.

Received 29 February; accepted 4 May 1988.

1. Akam, M. *Development* **101**, 1-22 (1987).
2. Morata, G. & Garcia-Bellido, A. *Wilhelm Roux Arch. dev. Biol.* **179**, 125-143 (1976).
3. Meinhardt, H. *J. theor. Biol.* **74**, 307-321 (1978).
4. Meinhardt, H. *Models of Biological Pattern Formation* (Academic, New York, 1982).
5. Bienz, M. *et al. Cell* **53**, (in the press).
6. Akam, M. E. & Martinez-Arias, A. *EMBO J.* **4**, 1689-1700 (1985).
7. Casanova, J., Sanchez-Herrero, E. & Morata, G. *Cell* **42**, 663-669 (1985).
8. Struhl, G. & White, R. A. H. *Cell* **43**, 507-519 (1985).
9. Ingham, P. & Martinez-Arias, A. *Nature* **324**, 592-597 (1986).
10. White, R. A. H. & Lehmann, R. *Cell* **47**, 311-321 (1986).
11. Martinez-Arias, A. & White, R. A. H. *Development* **102**, 325-338 (1988).
12. Bender, W. *et al. Science* **221**, 23-29 (1983).
13. Weinzierl, R., Axton, J. M., Ghysen, A. & Akam, M. *Genes Dev.* **1**, 386-397 (1987).
14. Casanova, J., Sanchez-Herrero, E., Busturia, A. & Morata, G. *EMBO J.* **6**, 3103-3109 (1987).
15. Campos-Ortega, J. & Hartenstein, V. *The Embryonic Development of* Drosophila melanogaster (Springer, Berlin , 1985).
16. White, R. A. H. & Wilcox, M. *Cell* **39**, 163-171 (1984).
17. Desplan, C., Theis, J. & O'Farrell, P. H. *Nature* **318**, 630-635 (1985).
18. Hiromi, Y. & Gehring, W. J. *Cell* **50**, 963-974 (1987).
19. Harding, K., Rushlow, C., Doyle, H. J., Hoey, T. & Levine, M. *Science* **223**, 953-959 (1986).
20. Hiromi, Y., Kuroiwa, A. & Gehring, W. J. *Cell* **43**, 603-613 (1985).

Erratum

A new superfamily of replicative proteins

T C Hodgman
Nature **333**, 22-23 (1988).

IN this piece of Scientific Correspondence, an unrevised version of the figure was used in error. The correct version is reproduced on the right.

```
Motif            I                                  II          III          IV                V                VI      Source
                                                                                                                          ref.
AlMV   821 VDGVAGCGKTTNIK                     55 RLIFDECFLQH 15 VIGFGDTEQIPF 22 ITWRSPADA  66 IFTTHE-AQGK-TFDNVYFCR 19 NGLVALSRH ( 1)
BMV    687 VDGVAGCGKTTAIK                     54 RLLVDEAGLLH 15 VLAFGDTEQISF 22 KTYRCPQDV  78 IKTVHE-AQGI-SVDNVTLVR 13 YCLVALTRH ( 1)
CMV    709 VDGVAGCGKTTAIK                     54 RVLVDEVVLLH 15 ALCFGDSEQIAF 22 TTFRSPQDV  79 IKTVHE-SQGI-SEDHVTLVR 13 YCLVAVTRH ( 1)
TMV    829 VDGVPGCGKTKEIL                     57 RLFIDEGLMLH 15 AYVYGDTQQIPY 24 TTLRCPADV  62 VHTVHE-VQGE-TYSDVSLVR 14 HVLVALSRH ( 1)
ToMV   829 VDGVPGCGKTKEIL                     57 RLFIDEGLMLH 15 AYVYGDTQQIPY 24 TTLRCPADV  62 VHTVHE-VQGE-TYADVSLVR 14 HVLVSLSRH (17)
TRV    901 VDGVPGCGKSTMIV                     56 VLHFDEALMAH 15 CICQGDQNQISF 24 ETYRSPADV  64 VSTVHE-SQGE-TFKDVVLVR 13 YLIVALSRH ( 1)
SFV    183 VFGVPGSGKSAIIK                     50 ILYVDEAFACH 16 VVLCGDPKQCGF 21 ISRRCTRPV  58 VMTAAA-SQGL-TRKGVYAVR 14 HVNVLLTRT ( 1)
SV     183 VIGTPGSGKSAIIK                     50 VLYVDEAFACH 16 VVLCGDPMQCGF 24 ISRRCTQPV  58 VMTAAA-SQGL-TRKGVYAVR 14 HVNVLLTRT ( 1)

IBV   1209 VQGPPGSGKSHFAI                     54 ILLVDEVSMLT 15 VVYVGDPAQLPA 30 KCYRCPKEI  82 VQTVDS-SQGS-EYDYVIFCV 11 RFNVALTRA (18)

BNYVV1 893 VKGGPGTGKSFLIR                     48 IIFVDEFTAYD 11 IYLVGDEQQTGI 25 MNFRNPVHD  72 KTTVRA-NQGS-TYDNVVLPV 12 LNLVALSRH (19)
BNYVV2 121 VLGAPGVGKSTSIK                     49 TMLVDEVTRVH 11 VICFGDPAQGLN 18 ASRRFGKAT  67 SILYSD-AHGQ-TYDVVTIIL 13 VRAVLLTRA (19)
BSMV2  267 ISGVPGSGKSIIVR                     41 LLIIDEYTLAE 11 VLLVGDVAQGKA 18 TTYRLGQET  62 CALAID-VQGK-EFDSVTLFL 12 LRLVALSRH (20)

uvrD    26 VLAGAGSGKTRVLV 17 MAVTF-TNKAAAEMRHRI 140 NILVDEFQNTN 16 VMIVGDDDQSIY 26 QNYRSTSNI 267 LMTLHS-AKGL-EFPQVFIVG 23 LAYVGVTRA (21)
rep     19 VLAGAGSGKTRVIT 17 AAVTF-TNKAAREMKERV 141 YLLVDEYQDTN 16 FTVVGDDDQSIY 26 QNYRSSGRI 271 LMTLHA-SKGL-EFPYVYMVG 22 LAYVGITRA ( 2)
recB    20 IEASAGTGKTFTIA 25 LVVTF-TEAATAELRGRI 303 VAMIDEFQDTD 18 LLLIGDPKQAIY 24 TNWRSAPGM 286 IVTIHK-SKGL-EYPLVWLPF 44 LLYVALTRS ( 4)
recD   164 ISGGPGTGKTTTVA 17 RLAAP-TGKAAARLTESL  48 VLVVDEASMID 16 VIFLGDRDQLAS 24 QLSRLTGTH 198 AMTVHK-SQGS-EFDHAALIL 11 LVYTAVTRA (22)

EBV     69 ITGTAGAGKSTSVS  7 CVITGTTVIAAQNLSAIL  88 VIVVDEAGTLS 26 IVCVGSPTQTDA 44 NNKRCTDVQ 426 AMTIAK-AQGL-SLNKVAICF  9 HVYVALSRA (23)
HCMV   117 VTGTAGAGKTSSIQ  7 CLVTGATTVAAQNLSQTL 100 IIVIDECGLML 26 IICVGSPTQTEA 44 NNKRCTDLD 513 AMTIAK-SQGL-SLEKVAVDF 10 HIYVAMSRV
HSV     94 ITGNAGSGKSTCVQ  7 CVVTGATRIAAQNMYAKL 111 VIVIDEAGLLG 26 LVCVGSPTQTAS 44 NNKRCVEHE 442 AMTITR-SQGL-SLDKVAICF  8 SAYVAMSRT ( 6)
VZV     87 ISGNAGSGKSTCIQ  7 CIITGSTRVAAQNVHAKL 110 VIVIDEAGLLG 26 IVCVGSPTQTDS 44 NNKRCQEDD 447 AMTIAR-SQGL-SLEKVAICF  8 SVYVAMSRT (24)

FIF    255 YTGSAGTGKSILLR  7 VAVTASTGLAACNIGGTI  21 ALVVDEISMLD 25 LIFCGDFFQLPP 29 KVFRQRGDV 219 MQTIHQNSAGKRRLPLVRFKA 33 QAYVALSRA ( 5)
             *  * ***        *   * *** **   *       ******        *****  *          *****          ***  ***  *   * * *        *******
Residue     V G AG GKS       VT  T  AA N    L       VDE           GD  Q          R            T   SQG  E  V              VALSR
distribn.   I A P    T       IA      T E    I       I                S                         L   AK   S  A              TLVT
            Y                A         R    V       F                                            VH   T                     GM
                                                                                                 NA   R                     SI
```

Creating hybridomas by electrofusion

M. Glassy

Electromanipulation has long been used to form pores in cell membranes to transfect them with foreign DNA. Now, the technique has been extended to the preparation of hybridomas.

THE availability of hybridomas and their products, monoclonal antibodies (MAbs), has revolutionized the healthcare industry. Virtually all modern biomedical laboratories utilize MAbs for diagnostic or research purposes, and therapeutic uses are under development. Most hybridomas are generated by fusing lymphocytes with antibody-secreting cancer cells using polyethylene glycol (PEG), which functions like a cationic glue to adhere the negatively-charged cell membranes to one another. Recently, a new technology for the production of hybridomas has been introduced: using precise, transient electric fields to induce somatic cell hybridization, a technique referred to as electrofusion[1-3].

The success of electrofusion depends upon both physical and biological parameters[2,3]. The physical variables include the voltage used to align, porate and fuse the cells, the design of the electrical generator, and the configuration of the chambers which house the cells. Biological parameters which may influence the outcome are the cell type, the fusion medium, and the genetic selection systems which allow non-fused cells to be sorted from fused cells.

Electrofusion is a three-step process. Cells are first exposed to an alternating electrical field (a.c.) to cause them to align into a "string of pearls" formation (Fig. 1) in preparation for fusion — a process named dielectrophoresis. The current is then switched to a direct electrical field (d.c.) to bring about the dielectric breakdown of the cell membranes to produce pores. The pores of adjacent cells form small channels between the two cells, which eventually broaden, causing the cells to fuse. When the direct electrical field is removed, the remaining pores in the membrane of the heterokaryon cell close, yielding an intact cell hybrid.

A shocking alternative

Electrofusion has been shown to be effective in the production of viable hybrids of plant[4,5], yeast[6], and mammalian[7-12] origin, including interkingdom hybrids[13]. The technique offers several distinct advantages over the use of PEG in the generation of hybridomas. It is less labour-intensive than the PEG-based procedure because it relies upon the biophysical effects of dielectrophoresis and membrane breakdown to induce heterokaryon formation, eliminating the need for a chemical incubation period and repeated washings. Because electrofusion depends on the physical effects of the electrical current, it functions regardless of the genetic or biochemical predispositions of the cells. Cell types which have been recalcitrant to PEG or other fusogens often can be incorporated into hybridomas using electrofusion. Moreover, electrofusion can be used with all cell types, including those such as human cell lines to which PEG is cytotoxic.

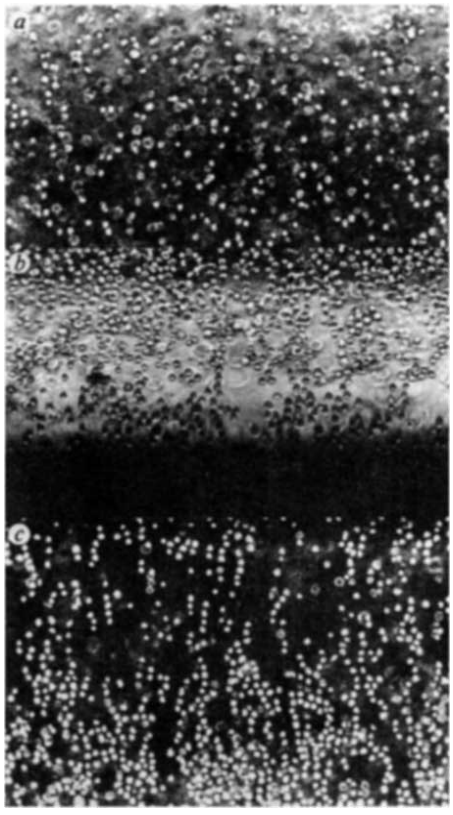


Fig. 1 Electrofusion sequence of UC 729-6 and human lymph node lymphocytes[12]. *a*, Time zero. *b*, After alignment, 20 sec at 580 V cm^{-1}. *c*, During the fusion sequence, 20 × 20 μsec at 7.69 kV cm^{-1}. The black area corresponds to one of the electrodes.

Cell preparation

A variety of options are available for the preparation of cells for electrofusion, both to maximize yield and to optimize the technique for the particular types of cells to be fused. Cells can be treated with facilitators, such as pronase, which cleave cell surface glycoproteins for more direct membrane-membrane interaction, to yield a larger number of hybrids. The fusion medium is usually an isotonic solution, but can vary from a non- to a low-electrolyte composition at physiological *p*H. The medium is typically 0.3 M mannitol, although other sugars have been used. The fusion of mammalian cells is facilitated by use of a 0.2-0.5 mM calcium acetate medium. Optimal cell concentrations vary according to the types of cells, but the ratio of isolated lymphocytes to fusion partners is generally between 2:1 and 4:1.

The type, style and geometry of the fusion chamber is critical. Whether a curved wire or a flat plate electrode is best for a particular fusion experiment is determined by the instrumentation and the number of times the fusion sequence must be repeated. At least one company which sells electrofusion instruments (BTX, Inc., San Diego, California) offers a database to customers which summarizes the optimal media, chamber configurations and other parameters for an experiment involving given fusion partners.

The amount of interest in electrofusion and the related field of electroporation is indicated by the number of recent meetings on the topic. An international workshop on electrofusion in hybridoma technology was held in Oslo, Norway in April, and in the same month two tutorials on the technique were held at the Federation of American Societies for Experimental Biology (FASEB) meeting in Las Vegas, Nevada. Scientists from around the world convened at the Oslo meeting, organized by Inger? Hagen of the Center for Industrial Research in Oslo, Norway, and Carl Borrebaeck of Lund University in Sweden, to discuss openly the advantages and the pitfalls of using electrofusion for the production of mouse and human MAbs. Topics discussed at the workshop ranged from the practical or "working man's" approach to human hybridoma generation, to microfusion with limited numbers of cells, the use of flow cytometry to optimize electrofusion parameters, avidin- biotin-mediated cell fusion, plant protoplast fusion, antibody engineering and electroporation-mediated gene transfer for cell immortalization

Application

To date, most of the experimental work in electrofusion has dealt with evaluating overall systems and verifying that the theoretical considerations do indeed

translate to practical use. But the field is still young, and electrofusion will not completely replace PEG fusions. For botanists, electrofusion is the technique of choice, but for the production of typical mouse hybridomas for MAb generation, PEG is still the easiest and most practical method. Electrofusion will have its greatest influence on the production of human MAbs. In many instances, for ethical and moral reasons, relatively few fusable human lymphocytes are available to investigators. Although it is very difficult to perform successful PEG fusions with less than 10^6 cells, electrofusion can theoretically be performed with as few as two cells, making it the more logical technique.

The area in which electrofusion could most be improved is the design of the cell fusion chamber. To allow the regulation of the experimental conditions in a timely and cost-effective manner, it would be very helpful to be able to observe the fusion events and readjust the parameters where appropriate. Most of the commercially available fusion chambers are housed in either metal casings or opaque materials that prevent the investigator from seeing the sequence of events as they occur. Glass slides are now on the market which allow direct observation, but their major flaw is volume: they typically hold less than 100 µl. This makes large-volume fusions particularly cumbersome, requiring time-consuming multiple repetitions.

Electrofusion offers a viable alternative for the production of hybridomas, but the potential of the technology has only begun to be explored. The major question to be answered now is whether the technique can compete with PEG in terms of ease of use and cost. For electrofusion to become widely used for the generation of hybridomas, the cost of the systems will have to be lower than current prices, and protocols will have to be developed that yield results superior to PEG. □

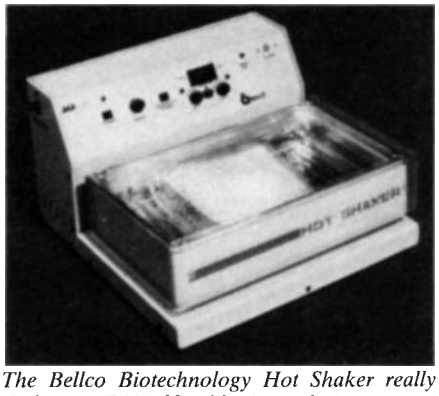
Mark C Glassy is at Biotherapeutics, Inc., 11025 N. Torrey Pines Road, La Jolla, CA 92037, USA. For more information, fill in reader service number 100.

1. Zimmermann, U. *Biochim. biophys. Acta* **694**, 227–277 (1982).
2. Hofmann, G.A. & Evans, G.A. *IEEE Engng med. Biol.* **5**, 6–25 (1986).
3. Jordan, C., Neumann, E., & Sowers, A. (eds) *Electroporation and Electrofusion in Cell Biology* (Plenum, New York in the press).
4. Bates, G.W., Gaynor, J.J., & Shekhawat, H.S. *Plant Physiol.* **72**, 1110–1113 (1983).
5. Saunders, J.A. *Front. Membr. Res. Agricult.* **9**, 147–156 (1985).
6. Schnettler, R., Zimmermann, U., & Emeis, C.C. *FEMS Microbiol. Lett.* **24**, 81–85 (1984).
7. Podesta, E.J. *et al. Eur. J. Biochem.* **145**, 329–332 (1984).
8. Scheurich, P. & Zimmermann, U. *Naturwissenschaften* **68**, 45–46 (1981).
9. Teissie, J. *et al. Science* **216**, 537–538 (1982).
10. Vienken, J. *et al. FEBS Lett.* **163**, 54–56 (1983).
11. Finaz, C., Leferve, A., & Teissie, J. *Exp. Cell Res.* **150**, 477–482 (1984).
12. Pratt, M., Mikhalev, A., & Glassy, M.C *Hybridoma* **6**, 469–477 (1987).
13. Cocking, E.C. *Ciba Found. Symp.* **103**, 119–128 (1984).

Tissue culture tips

Techniques to tease tissues into growing in non-native environs will be the focus of the Tissue Culture Association meeting next week in Las Vegas, Nevada. A preview of the exhibits is below.

BELLCO Biotechnology sells a "hybrid" instrument for carrying out **nucleic acid hybridizations** which is shaking up rDNA filter-blotting techniques: the Hot Shaker (*Reader Service No. 101*). Bellco's $1,575 (US) Hot Shaker combines temperature control to within one-tenth of a degree Celsius and agitation of up to 60 oscillations per minute into a single piece of laboratory equipment. The water bath holds standard-sized baking dishes for gel or membrane work, and has a lid to prevent bath evaporation. The heater has a 105 °C maximum safety thermostat which sets off an audible alarm if the temperature rises. A d.c. output allows the Hot Shaker to be connected to an optional chart recorder for printing out results. Bellco Glass will be exhibiting in booth 501.



The Bellco Biotechnology Hot Shaker really shakes up rDNA filter blotting techniques.

One of the latest products from BRL, exhibiting in booths 411 and 412, is the KiloBase **DNA sequencing system** (*Reader Service No. 102*). BRL says the KiloBase system produces accurate sequences of up to 1,000 bases per reaction set using M13 templates. If KiloBase does not provide researchers with more sequence data than the system they are currently using, BRL offers a money-back guarantee. The KiloBase system uses a modification of the Sanger chain-termination technique. It works with either ^{35}S- or ^{32}P-labelled nucleotides and either single- or double-stranded DNA templates. The enzyme used is the highly characterized Klenow fragment of DNA polymerase I — an enzyme capable of "proofreading" the sequence synthesized and removing nucleotides that have been incorrectly incorporated.

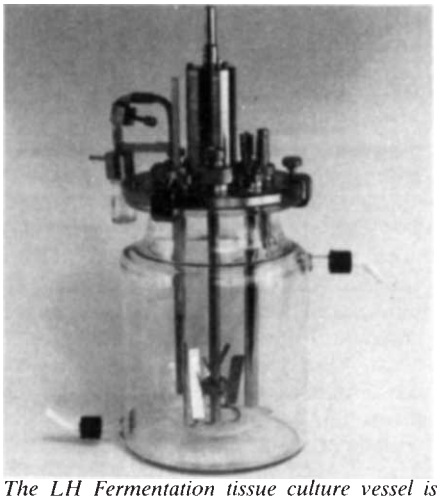
BRL's KiloBase sequencing system.

Suitable environments

A **rotary drum mixer** that can be used for tissue culture, cytotoxicity assays and the growth of viruses is available from Wheaton Instruments, exhibiting in booth 303 (*Reader Service No. 103*). The drum mixer comes with an 8-inch in diameter drum made from electroplated stainless steel, that can hold 18 small bottles or vials up to 32 mm in diameter. Because the mixer's circuit board is coated with a protective sealant, it can be used in an incubator. The tilt angle of the mixer can be adjusted from 0° to 90°, and the speed range is 1–25 r.p.m. Wheaton sells the rotary drum mixer for $346 (US).

Wheaton's wheel goes round and round at just the right speed for tissue culture.

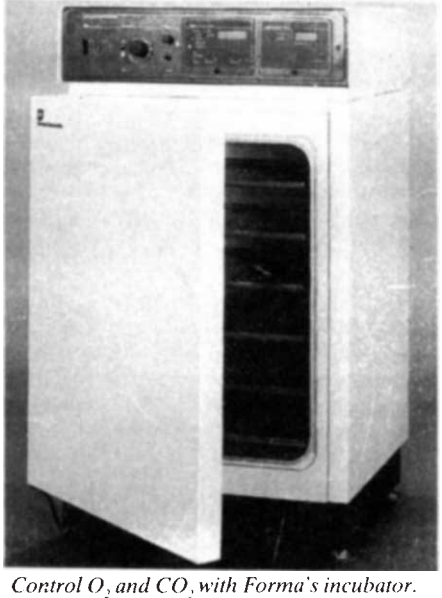
LH Fermentation, in booth 304, has extended its range of benchtop fermenters to include **tissue culture vessels** in response



The LH Fermentation tissue culture vessel is easy on suspended or microcarrier-bound cells.

to the growing demand for fermentation systems suitable for animal cell culture (*Reader Service No. 104*). The company says the improved impeller design, domed fermenter base and low agitation speeds minimize the shear forces inside its 500 series tissue culture vessels, making them appropriate for suspension cultures and microcarrier-based cultures. Stirrer types include pitch paddle and marine and Rushton impellers, all of which are height-adjustable and have speeds covering the range of 0–250 r.p.m. The two available 1.3-and 2.3-litre vessels have 13 ports and are fully jacketed to allow accurate temperature control and prevent localized overheating at low agitation rates. The 500 series systems are complemented by a range of interchangeable control modules, electrodes and other components.

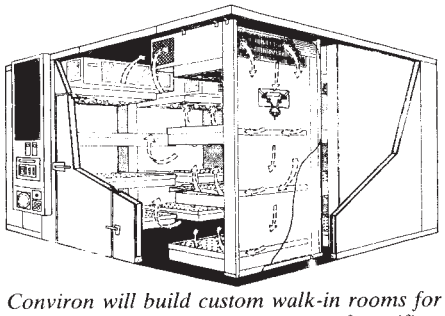
Forma Scientific, in booth 203, has introduced a new **cell culture incubator** designed for the automatic control of O_2 and CO_2 (*Reader Service No. 105*). Using electronic O_2 and CO_2 sensors, the $3,595 (US) model 3159 incubator maintains



Control O_2 and CO_2 with Forma's incubator.

programmable O_2 and CO_2 levels *in vitro* to simulate a wide range of *in vivo* conditions for the optimal growth of mammalian cells. Forma Scientific says the O_2 sensor technology supersedes constant flow multiplexed and wet chemistry methods, and automatically calibrates. Gas inject indicators and alarm systems for O_2, CO_2 and temperature are standard. Other features include water-jacketed construction, self-diagnostic electronic control systems, gas sampling ports and an access port for external probes and electrical lines.

Conviron, exhibiting in booth 4, has recently developed **walk-in rooms** for plant tissue culture that it says eliminate temperature uniformity and gradient problems (*Reader Service No. 106*). The rooms incorporate individual insulated tissue culture modules, which are installed along one or both side walls according to customer needs. Each module contains



Conviron will build custom walk-in rooms for plant tissue culture to a wide range of specifications.

two or three air shelves measuring 30 in. × 48 in., with fluorescent lamps attached underneath to illuminate the shelf below. Up to six modules — three along each side wall — may be provided to yield up to 180 square feet of illuminated shelf space. The temperature of the room can be varied between 15 °C to 40 °C with the lights on; with the lights off, the temperature can be lowered to 5 °C. Conviron will build rooms ranging from 9 × 6 feet to 11 × 16 feet. All rooms include a microprocessor control system to allow step or ramp programming of temperature, humidity and light intensity.

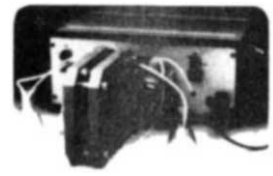
In booth 503, Boehringer Mannheim Biochemicals will be exhibiting its specialist range of **reagents for cell culture** (*Reader Service No. 107*). Intended particularly for the biotechnologist who has a requirement for chemically defined serum-free media, the range includes media supplements, growth factors, and enzymes for tissue dissociation and removal from culture vessels. Media supplements include the Nutridoma range for the growth of hybridomas in low protein conditions for easier isolation of monoclonal antibodies. Albumin, transferrin and soybean lipids are supplied for use in Iscove media. Boehringer Mannheim Biochemicals also sells interleukin-1 and interleukin-2; endothelial cell, epidermal cell and fibroblast growth factor; insulin-like growth factor; nerve growth factor; and platelet-derived growth factor.

For keeping cultures in the right climate, Becton Dickinson Microbiology Systems has added new items to its line of **disposable atmospheric generation pouches** (*Reader Service No. 108*). To its GasPak Pouch, Campy Pouch and GasPak Sak, Becton Dickinson has added the GasPak CO_2 pouch. The GasPak CO_2 pouch produces a 3–12 per cent CO_2-enriched atmosphere which is conducive to the cultivation of capnophilic organ-

ADVERTISEMENTS

LIPOSOMES NEW

REPRODUCIBLE — HOMOGENEOUS — EXTREMELY RAPID



Our new LIPOSOMAT guarantees rapid (5ml within 60 minutes) preparation of uniformly sized unilamellar liposomes. Liposome size is selectable between 25 and ca. 600 nm diameter. Sample volume is normally 5-10ml or up to 200ml if used in combination with an additional instrument. Lipid concentration can be up to 300 mg/ml.

BIANORM

POB 126, D-8000 Munich 65, FRG

Reader Service No.48

SINGER INSTRUMENTS

micro injection?
micro dissection?
micro surgery?
micro positioning?

Micromanipulators

Singer Instrument Co. Ltd.

Trebotough Lodge, Roadwater, Watchet, Somerset, England, TA23 0QL. Tel: Washford (0984) 40226
Telex: 337492 Comcab G

Reader Service No.50

Copies of articles from this publication are now available from the UMI Article Clearinghouse.

Mail to: University Microfilms International, 300 North Zeeb Road, Box 91, Ann Arbor, MI 48106

Becton Dickinson puts microbes in the bag with its controlled environment pouches.

isms such as *Neisseria gonorrhoeae, N. meningitidis,* and *Haemophilus influenzae*. Each pouch holds up to three 100-mm Petri dishes and is closed by a reusable sealing bar. The pouches do not require the crushing of glass vials, and work without heat generation. The GasPak CO_2 pouches cost $39.85 (US) for a box of 50 incubation pouches with integral reagent sachets, and 50 liquid activating reagent packets. Becton Dickinson is in booth 601.

Containers for growth

Nalge, in booth 401, will have available copies of its **1988 Nalgene labware catalogue** (*Reader Service No. 109*). The 150-page catalogue details the full line of over 450 Nalgene plastic labware products. Full product specifications, prices, data on chemical resistance and physical properties, and care and use instructions are included. New products include 30- and 60-ml Nalgene square bottles, 4-mm syringe filters, a water quality membrane filter, a Teflon PFA evaporating dish, a shelf rack for plastic animal cages, and a refrigerated collection rack for metabolic animal cages. Other new products are a large beta waste shield, a leakproof bottle mailer, and a reusable aerosol spray bottle for thin-layer chromatography applications. New to the cryoware product range are a transparent tube and sleeve for cryovials, and colour-coded cryovial tabs.

Elkay Products, exhibiting in booth 3, has just introduced a **custom printing service** for plastic laboratory consumables (*Reader Service No. 110*). Elkay will print cylindrical products from 0.5 in. to 6 in. in diameter and up to 9 in. in length, including test tubes, centrifuge tubes, transport and mailing tubes and specimen containers. Any of the over 500 PMS colours can be used, allowing researchers the versatility of colour-coding any project. A variety of ink types are available, including write-on and autoclavable inks. The colour-coded consumables can be used to colour-code diagnostic kits to identify hazardous samples, to personalize laboratory sample tubes with a company name, and to highlight volume graduations.

Your company name on consumables from Elkay.

Maximizing yield

Coulter has a system that it says provides exact clone selection for total purity in **monoclonal antibody production** — the Auto-Clone system (*Reader Service No. 111*). When used in conjunction with Coulter's EPICS flow cytometer, it provides a fully programmable, automated system for the deposition of single cells into multiwell plates for specifically selected cultures. After sorting by the flow cytometer, cells are placed into the individual wells of 24-, 60- or 96-well microculture plates in multiples of 1 to 10, or 100 to 1,000. The deposition sequence is microprocessor-controlled; commands are entered through a front panel keyboard. Coulter, exhibiting in booth 707, sells the Auto-Clone for £7,500 (UK).

The **cell culture qualification kit** from Miles Diagnostics provides a do-it-yourself method for determining cell requirements for growth and proliferation (*Reader Service No. 112*). The starter six-pack kit includes ingredients known to be essential for cell growth: transport proteins such as transferrin and albumin, hormones such as insulin, sodium selenite, lipids, and fibronectin, an attachment factor. The qualification kit contains enough material for the preparation of up to 10 litres of serum-free media. Miles Diagnostics is in booth 103.

In addition to displaying its Opticell culture system, Charles River Biotechnical Services will be offering details on its ***in vivo* contract production services** in booths 104 and 105 (*Reader Service No. 113*). Charles River says it is the largest contract producer of monoclonal antibody: last year the company reproduced nearly 5 kg of antibody *in vivo*. Its affiliation with Charles River Laboratories allows it to produce murine-derived hybridoma cell lines in mouse peritoneal ascites fluid to achieve its high productivity levels. Charles River Biotechnical Services also provides hybridoma assessment and evaluation services to determine the feasibility of using its ascites system for the large-scale production of monoclonal antibody from clients' cell lines. The package includes cell line quality assurance testing for murine viral contamination, sterility and mycoplasma detection. The company will provide 100 mg of antibody for client analysis, and will produce a bulk sample of ascites fluid from 200 mice.

Analysis

Flow Laboratories, in booth 208, has a **fluorometer** which reads samples in a microplate format — the Fluoroskan II (*Reader Service No. 114*). The $22,500 (US) Fluoroskan II permits users who require the sensitivity of fluorometry to carry out their assays in microplates. The Fluoroskan II has four excitation and four emission filters which can be selected automatically in any combination from the membrane-covered keypad. The instrument will blank on either a well, column or row of a microplate, permitting flexible assay formatting. Results may be expressed in either absolute fluorescence units or relative to a user-selected standard arbitrarily defined as 100 per cent. The bidirectional RS232 interface of the Fluoroskan II allows it to send results to a computer, and allows the computer to control its functions. The thermoblock plastic of the Fluoroskan II's plate carriage can be heated automatically to a stable 37 °C for assays which require the establishment of an elevated thermal equilibrium.

YSI, in booth 607, sells an **automated glucose and L-lactate analyser** for the determination of glucose and L-lactate levels in fermentation broths and cell cultures (*Reader Service No. 115*). The $6,600 (US) model 2000 uses the same enzyme electrodes incorporated into the YSI industrial analyser.

YSI's tool for analysing fermenter broths.

The system measures both sugars simultaneously, and prints and displays the results in 60 seconds. The instrument automatically calibrates, and can run up to 60 samples per hour, says YSI. The model 2000 can quantify glucose in concentrations of up to 20 gl^{-1}, and L-lactate in concentrations of up to 2 gl^{-1}. Sample analysis requires little or no preparation: the sipping tube automatically aspirates fluid for analysis. An RS-232C port is provided for connection to a computer. □

These notes are compiled by Carol Ezzell from information provided by the manufacturers. To obtain further details about these products, use the reader service card bound inside the journal. Prices quoted are sometimes nominal, and apply only within the country indicated.